Biodiversity conservation and evolutionary models

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Doctor of Philosophy
in the
University of Canterbury
by
Klaas Hartmann

Examining Committee

Prof. Vincent Moulton	External (overseas) examiner
Prof. Allen Rodrigo	External (NZ) examiner
Prof. Mike Steel	Internal examiner

University of Canterbury
2008

For Ullr, Skaði, Cabrakan and the Naiads for some excellent times and Noah – one of the first conservationists!

Abstract

Biodiversity conservation requires a framework for prioritising limited resources to the many endangered species. One such framework that has seen much attention and is considered extensively in this thesis, is the Noah's Ark Problem (NAP). The NAP combines a biodiversity measure (Phylogenetic Diversity; PD) with species survival probabilities and conservation costs. The aim of the NAP is to allocate the limited conservation resources such that the future expected PD is maximised.

Obtaining optimal solutions to the NAP is a computationally complex problem to which several efficient algorithms are provided here. An extension to the NAP is also developed which allows uncertainty about the survival probability estimates to be included. Using this extension we show that the NAP is robust to uncertainty in these parameters and that even very poor estimates are beneficial. To justify using or promoting PD, it must produce a significant increase in the amount of biodiversity that is preserved. We show that the increase attainable from the NAP is typically around 20% but may be as high as 150%.

An alternative approach to PD and the NAP is to prioritise species using simple species specific indices. The benefit of these indices is that they are easy to calculate, explain and integrate into existing management frameworks. Here we investigate the use of such indices and show that they provide between 60% and 80% of the gains obtainable using PD.

To explore the expected behaviours of conservation methods (such as the NAP) a distribution of phylogenetics trees is required. Evolutionary models describe the diversification process by which a single species gives rise to multiple species. Such models induce a probability distribution on trees and can therefore be used to investigate the expected behaviour of conservation methods. Even simple and widely used models, such as the Yule model, remain poorly understood. In this thesis we present some new analytic results and methods for sampling trees from a broad range of evolutionary models. Lastly we introduce a new model that provides a simple biological explanation for a long standing discrepancy between models and trees derived from real data – the tree balance distribution.

Table of Contents

List of Figures

Acknowledgments

First and foremost I would like to thank my supervisor Mike Steel for his support, guidance and encouragement. Mike started me on a very enjoyable and rewarding research project that I have enjoyed tremendously. Secondly I would like to thank Arne Mooers. Arne provided great support and with his help I have tried to make my work as applicable and relevant as possible.

Throughout this work I have had many fruitful discussions with people from all over the world. In particular I would like to thank Tanja Gernhard. It took us a while to recognise the similarities in our work, but when we did it proved most productive. I would also like to thank Paul Armsworth, Richard Malowney, Vincent Moulton, Fabio Pardi, David Redding, Sébastien Rioux Paquette, Charles Semple, Tobias Thierer, Rutger Vos, Margee Will, Oliver Will and Dennis Wong for the interesting discussions and support that they have given me during my PhD.

A number of institutions supported me financially during my work for which I am very grateful. First and foremost the "University of Canterbury" and the "Allan Wilson Centre for Molecular Ecology and Evolution" provided a generous scholarship. Google Inc. and NESCENT supported me through a Summer of Coding scholarship which enabled me to implement some of my algorithms. Arne Mooers at Simon Fraser University and IRMACS supported me during my lengthy visit to Vancouver and the Isaac Newton Institute hosted me and funded some of my travel expenses whilst at Cambridge for the Phylogenetics program.

My fellow PhD students provided great social times with many informal talks, games nights, Vic & Whale sessions, skiing and pub quizzes throughout my time in Christchurch. Kristine was wonderfully supportive, even during my grumpiest, "I'm sick of my thesis" times and proof-read this entire manuscript. Despite understanding little of the mathematical details, Krissy's proof-reading was invaluable, even uncovering a mistake in a published paper... Lastly I would like to thank Damian and Rebecca – our flatmates during the last months of my PhD – for putting up with my anti-social thesis writing behaviour during this time!

Preface

This thesis has been divided in two parts, that tackle very different but yet inextricably linked topics. The first part deals with evolutionary models. Despite their wide use the mathematical properties of evolutionary models are relatively poorly understood. In this part we derive new analytic properties of existing models and provide some simple algorithms for sampling trees from arbitrarily complex models. Lastly we introduce a new model that provides a better fit to existing data sets than the current widely used models.

The second part of this thesis considers biodiversity conservation problems. In particular if we have a limited budget to allocate to the conservation of species, how should this be allocated to ensure that as much biodiversity as possible is retained? In this part we, provide new algorithms for solving existing conservation frameworks, address uncertainty in the input parameters for these frameworks and contrast the different approaches. We use the results from the first part of the thesis to consider the expected behaviour of these frameworks for trees produced by evolutionary models. This permits us to investigate the expected behaviour in real situations.

Four publications from this thesis have been published or accepted, another publication has been submitted and three further papers are in preparation. Due to the interdisciplinary nature of this work, much collaboration was involved – in total there are ten authors on the publications resulting from this thesis (see Table 1). It should be noted that I was lead author or equal contributor on all of these papers. Various chapters draw heavily on these publications, however the work presented here is either my own work or work to which I contributed substantially. A couple of results from these publications that were not my own work have been reproduced in succinct form for completeness and these results have been clearly attributed. Notably, (i) the approaches in sections 2.3 and 3.2 were developed by Tanja Gernhard, I implemented these method and helped clarify the text, (ii) The

Chapter	Reference	Journal	Status
2,6,9	Gernhard, Hartmann and Steel	J. Math. Biol.	in press
3	Hartmann et al. (a)	Syst. Biol.	submitted
4	Hartmann et al. (c)		in prep.
Part II Intro., 6	Hartmann and Steel (2007)	Book chapter	published
7	Hartmann and Steel (2006)	Syst. Biol.	published
8	Hartmann et al. (b)		in prep.
9	Hartmann and Mooers		in prep.
10	Redding et al. (2008)	J. Theor. Biol.	in press

Table 1: A summary of the publications produced from this thesis and the chapters which utilise the results in these publications. Klaas Hartmann is lead author or contributed equally to all the publications listed here.

first part of section 6.2 (prior to section 6.2.1) is a result from Steel (2006), that has been included here for contrast with my own work for a related process.

Notation

Here we briefly introduce some of the notation used throughout this thesis, this notation is explained in further detail where it is first used. We let $\mathcal{T}$ denote a phylogenetic tree, that is, a tree whose leaves comprise the set of taxa (generally species or populations) under study, and whose remaining vertices (nodes) are of degree at least 3 (the degree of a vertex is the number of edges that are incident with it). The vertices at the tips are called *leaves*. If all the non-leaf vertices in a tree have three incident edges the tree is said to be *fully resolved* (sometimes called 'binary' - these are the trees without polytomies, and so are maximally informative).

Trees may also have some ancestral vertex of degree one or two distinguished as a root vertex. The root vertex or simply root, is the ancestral species from which all species in the tree are descendant. There is therefore a natural direction of time from the root to the leaves of the tree. As illustrated in Figure 1 the root vertex may or may not have an associated edge length depending on the context of the work. Most of the work in this thesis deals with *rooted binary trees*.

The length of an edge, i, in the tree is denoted by λ_i, for rooted trees the

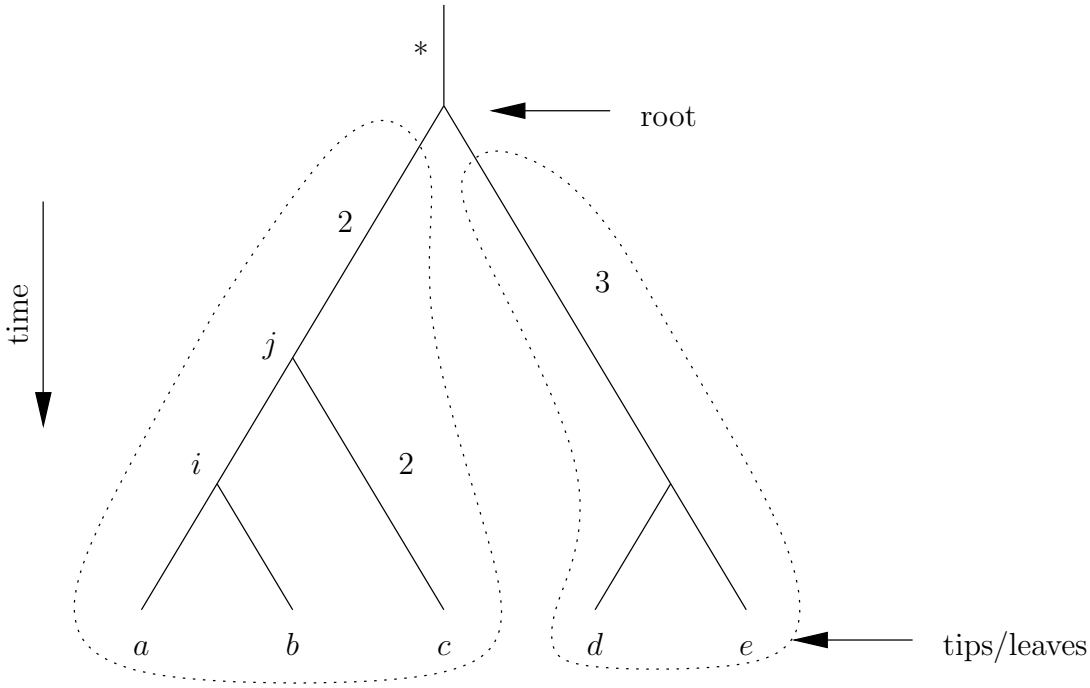


Figure 1: This figure shows an example of a rooted binary tree. Time progresses down the page from the root to the tips/leaves. All edges have length one unless otherwise indicated. Depending on the context the root may have an edge (indicated by $*$) associated with it. The two trees descended from the root are circled, we refer to these as $\mathcal{T}_a$ and $\mathcal{T}_b$, or τ_a and τ_b. The children of a node are the leaves below it such that we have: $\mathcal{C}_i = \{a, b\},\ C_i = 2$ and $\mathcal{C}_j = \{a, b, c\},\ C_j = 3$. Lastly, to illustrate edge lengths we have $\lambda_i = 1$ and $\lambda_j = 2$.

length of the edge ancestral to a species j is denoted by λ_j (there is a one-to-one correspondence between nodes and edges). We consider a phylogenetic tree, $\mathcal{T}$, to include both the tree structure and the corresponding set of edge lengths – $\lambda_{\mathcal{T}}$. If the edge lengths and leaf labels are omitted we refer to the tree as a tree shape and denote it by τ. For brevity $|\mathcal{T}|$ and $|\tau|$ are used to denote the number of species in a tree.

Lastly, for rooted trees, the set of species below a node i (separated from the rest of the tree by i) are referred to as the children of i. The set of these nodes is denoted by $\mathcal{C}_i$ and the size of this set by C_i. A summary of symbols used throughout this thesis (including concepts not addressed here) is given in Table 2 and some of these are illustrated in Figure 1.

Symbol	Meaning
$\mathcal{T}$	Phylogenetic tree with edge lengths
$\mathcal{T}_a$,$\mathcal{T}_b$	The two trees descendant from the root of $\mathcal{T}$
λ_i	Length of edge i or of the edge above vertex i
$\lambda_{\mathcal{T}}$	Set of all edge lengths of $\mathcal{T}$
τ	Phylogenetic tree without edges or leaf labels
$\lvert\mathcal{T}\rvert$, $\lvert\tau\rvert$	The number of species in the tree
$\mathcal{C}_i$	The set of vertices below edge or vertex i
C_i	The number of vertices below edge or vertex i ($C_i = \lvert\mathcal{C}_i\rvert$)
$\mathring{V}$	The set of interior vertices of a tree
a_i	The survival probability of species i if it is not conserved
b_i	The survival probability of species i if it is conserved
c_i	The cost of conserving species i
β	Speciation rate
μ	Extinction rate
$g(u)$	Probability density of the time to the next speciation event (u)

Table 2: Some of the notation used throughout this thesis is listed here. Further explanations are given as the symbols are introduced.

Part I

Evolutionary Models

Chapter I

Introduction

In the first part of this thesis we consider evolutionary models and produce results that are used in the second part to explore biodiversity conservation problems. Evolutionary models describe the process by which a single ancestral species diversifies into the species that are extant today. In addition to speciation events, these models may include other features of evolution such as extinction and trait evolution.

Evolutionary models have been developed for many reasons. One of their main uses has been to try to explain the evolution of biological diversity for organisms. Studies in this field try to fit a developed model to a data set (a record of fossil presence through time or a phylogeny). Fitting models of evolution to a data set is an important part of hypothesis testing, and an integral part to the scientific method (for example studies such as Sepkoski (1982), Bininda-Edmonds et al. (2007); and for a review Mooers and Heard (1997), Mooers et al. (2007)).

Another use of evolutionary models is to explore the expected behaviour of an algorithm or method. For example consider a method that performs poorly on some trees; whether this is problematic depends on the likelihood of those trees occurring. This can be investigated by running the method on a sample of trees from a realistic evolutionary model and analysing its behaviour for those trees. This type of approach is utilised in Part II of this thesis to investigate the expected performance of a range of biodiversity conservation approaches.

1.1 Simple evolutionary models

Here we provide a brief overview of two common evolutionary models that will be used throughout this thesis. Arguably the simplest evolutionary null model is the Yule model (Yule, 1924; Harding, 1971). Under the Yule model each species has an equal probability of undergoing a speciation event at any given point in time. The time between speciation events on a lineage is therefore exponentially distributed with parameter β. The Yule model has been widely used as a null model with which to compare real phylogenetic trees and explore evolutionary hypotheses (Aldous, 2001; Mooers and Heard, 1997). Despite its wide application, the Yule model has many mathematical aspects that remain uncharacterised.

The Yule model does not include explicit extinction events. Extinction is generally considered to be included implicitly by treating the speciation rate as a net speciation rate. If this approach is taken the speciation rate should change over time as discussed in section 4.3.2, however for the Yule model the speciation rate is independent of time.

The constant rate birth-death model (Mooers and Heard, 1997) is closely related to the Yule model and features explicit extinction events. Extinction events are modelled in a similar manner to speciation events – there is a constant extinction rate for each species. The time to an extinction event on a lineage (provided a speciation event does not occur) is therefore exponentially distributed with parameter μ.

A model with explicit extinction events will produce a tree like the left tree in Figure 1.1. This tree will include lineages that have become extinct. If a tree is reconstructed from 'real data' the sequences available will usually correspond to modern extant species, hence we will have no knowledge of the extinct lineages. The right tree in Figure 1.1 shows the tree we would hope to construct from the extant species in the left tree. Throughout this thesis we generally assume that the tree we are dealing with corresponds to a reconstructed tree containing only extant species.

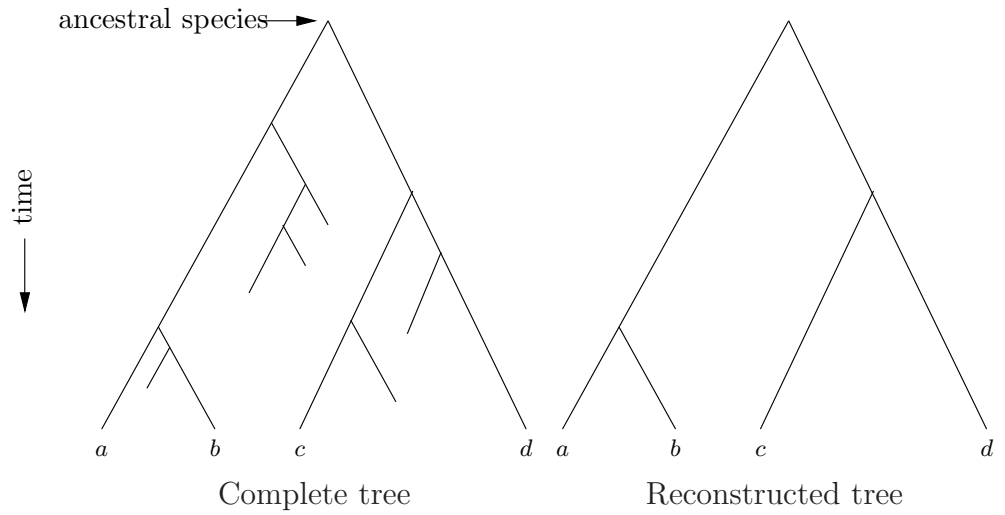


Figure 1.1: The left tree shows the full evolutionary history of all the species descendant from the single ancestral species. This includes extinct lineages which are readily identified as those that end prematurely. The right tree shows the tree that could be constructed from the modern extant species (a, b, c and d) thereby excluding the extinct species.

1.2 Part I overview

Despite the widespread usage of evolutionary models, even simple models such as the Yule model are still poorly understood. In chapter 2 we produce a mathematical framework for calculating probability densities associated with a class of evolutionary models, including the Yule model. For a given evolutionary model in this class the framework permits us to find (i) the probability density of the tree shape, (ii) of any edge length for a given tree shape and (iii) of the longest pendant edge length for a given tree shape.

For complicated models analytic solutions may be difficult to obtain, consequently model characteristics are frequently explored using sampled trees. Despite the widespread use of this approach, tree sampling is poorly understood. In chapter 3 we discuss a widely used program for sampling trees – Phylogen (Rambaut, 2002) – that has been used inappropriately in many studies. We then introduce alternative sampling methods that are simple and can be applied to most evolutionary models. These methods (to be published in Hartmann et al. (a)) are the first published methods for sampling

phylogenetic trees from evolutionary models of which we are aware.

For over a decade it has been widely known that phylogenetic trees produced by simple evolutionary models differ from those constructed from 'real' data (Mooers and Heard, 1997; Aldous, 2001). Despite this knowledge no simple evolutionary model with a satisfying biological explanation has been published. In chapter 4 we present a new evolutionary model that matches trees constructed from 'real' data. This model also has a simple biological explanation – after a speciation event the new species have a heightened probability of speciating. This model is compared with trees from several large databases to which it gives a good fit. This provides one possible answer to a long standing problem of evolutionary models.

Chapter II

Stochastic properties of generalised Yule models

In this chapter two approaches are presented for calculating edge length probability distributions for Yule models. This chapter is contained in Gernhard, Hartmann and Steel (some derivations in that paper that were solely the work of Tanja Gernhard have been omitted here).

The first approach we present applies to a more general class of evolutionary models based on Bellman-Harris (BH) processes which we describe here and refer to as BH models. For trees produced by the BH model we provide methods for calculating (i) the probability density of the tree shape, (ii) of any edge length for a given tree shape and (iii) of the longest pendant edge length for a given tree shape. For Yule models, analytic solutions are obtainable, however for some BH models it may be necessary to solve the required integrals numerically. These methods extend the results in Steel and McKenzie (2001) and have been applied in Redding et al. (2008). Other related properties can be readily explored by extending our methodology.

The second approach we present utilises rank functions to obtain edge length probability densities for the Yule model. This approach was first introduced in Gernhard et al. (2006) for expectations; here it is extended to give distributions and to permit a known age of the tree to be incorporated.

Our methods can be useful in many contexts including testing evolutionary hypotheses, constructing phylogenetic trees, and biodiversity conservation. In subsequent chapters we apply our methods to biodiversity conservation problems (chapters 8 to 10) and use them to investigate an alternative speciation model (chapter 4).

2.1 Evolutionary models

Throughout this chapter, we consider rooted binary trees. The root represents the ancestral species from which all other species are descendant. Internal nodes (with degree three) are ancestral species and the leaves (nodes of degree one) correspond to their modern descendants. The edges between any two nodes have associated lengths which may be interpreted as the time between speciation events or the genetic difference between the species corresponding to those nodes; this interpretation will depend on the data from which the tree was derived.

The Yule model makes the simple assertion that each species is equally likely to undergo a speciation event at any given point in time. Speciation can therefore be considered as a Poisson process on any given lineage and the time between speciation events on a lineage is exponentially distributed with rate β, in various examples throughout this chapter we set $\beta = 1$. In a Bellman-Harris (BH) process an individual has a random lifespan, u, described by a probability distribution, $g(u)$, after which the individual is replaced by a random number of new individuals. Note that every species speciates according to the same distribution g. The Yule model is therefore analogous to a BH process where the 'lifespan' of an individual is the time between speciation events on a given lineage (which is exponentially distributed) and each species is replaced by two new species (only binary trees are considered here).

This connection between BH processes and the Yule model suggests that it may be worthwhile to consider the larger class of BH evolutionary models. The BH models proposed here retain the constraint that each species is replaced by two new species, however the time between speciation events on a given lineage may be distributed according to an arbitrary probability density, $g(u)$. BH processes have been considered extensively in the mathematical branching process literature (Jagers, 1975; Sankaranarayanan, 1989), particularly as applied to birth and death processes, however they have seen little application to phylogenetic trees (Aldous, 2001).

The motivation for introducing BH models here is simply that our first method applies to the entire class of BH models. As such we do not discuss

the implications of BH models further (or investigate different probability densities for $g(u)$). It should be noted that analytic solutions for all the applications presented here exist for the Yule model but no such guarantee exists for other BH models. Solutions for these models may need to be found numerically, which introduces additional complications due to the nested nature of some of the integrals.

The BH model we consider is restricted to binary trees. Approach 1 can readily be adapted to multifurcations, however the biological motivation for such processes seems limited. An interesting extension would be to consider the more general Crump-Mode-Jagers models Crump and Mode (1968, 1969) which would allow a speciation to occur without replacing the original species.

2.2 Approach 1: Using nested integrals

The first approach describes the probability of a tree recursively in the form of nested integrals. These integrals will be nested to the same order as the depth of the tree. Our method applies to all BH models, however for some models the integrals may need to be solved numerically. Fortunately for the Yule model we can show that analytic solutions to these integrals exist.

2.2.1 Probability density of a BH-tree

Let τ denote the shape of the a tree, that is the tree without the associated edge lengths (see Figure 2.1). The two trees descendant from the root of τ are denoted by τ_a and τ_b; the number of species (leaves) in a tree is given by $|\tau|$. A tree, τ, may have edge lengths associated with it; the set of all edge lengths is denoted by λ_τ and the length of an individual edge, e, is denoted by λ_e. The root edge is denoted by r and its descendants are a and b, thus their edge lengths are λ_r, λ_a and λ_b respectively. For BH models the distance between the root node and any leaf node is the same for all leaves (the tree is ultrametric) and is denoted by t.

Using this notation the probability density for a tree, τ, with specified edge lengths under a BH model can be stated recursively as the product of the probability density of the root edge and the probability density for the trees descendant from the root:

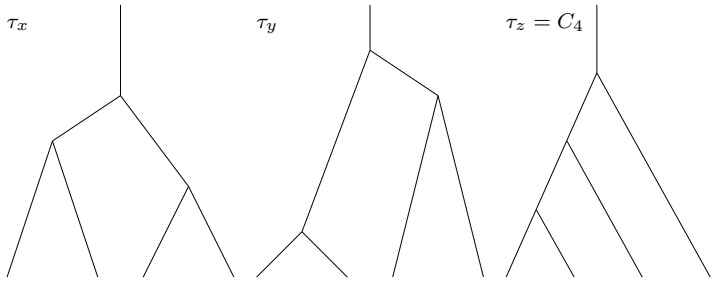


Figure 2.1: Two trees have the same shape if they are indistinguishable when branch lengths and leaf labels are disregarded, thus the two trees on the left (τ_x and τ_y) have the same shape whereas the tree on the right (τ_z) has a different shape. The latter is an example of a caterpillar tree (C_n, $n = 4$) where each internal node has one species as a direct descendant.

$$h(\tau, \lambda_\tau) = \begin{cases} \nu_\lambda g(\lambda_r) h(\tau_a, \lambda_{\tau_a}) h(\tau_b, \lambda_{\tau_b}) & |\tau| > 1 \\ 1 - \int_0^{\lambda_r} g(u) du & |\tau| = 1, \end{cases}$$

where ν_λ equals two if $\lambda_a \neq \lambda_b$ and one otherwise. This factor of two arises as the side on which each descendant tree occurs is irrelevant. If the tree is of size one, the probability of obtaining it is simply one minus the probability of a speciation event occurring too soon (resulting in a tree with more than one species). Note that the edge lengths are continuous variables, hence $h(\tau, \lambda_\tau)$ is a probability density.

The shape of a tree depends only on the number of species descendant from each internal node, thus two trees have the same shape if they are indistinguishable after the edge lengths and leaf labels are disregarded (see Figure 2.1 for an example). The probability of obtaining a particular tree shape after time t from a BH model can also be stated recursively (see also Steel and McKenzie (2001)) :

$$p(\tau|t) = \begin{cases} \nu_\tau \int_0^t g(\lambda_r) p(\tau_a | t - \lambda_r) p(\tau_b | t - \lambda_r) d\lambda_r & |\tau| > 1 \\ 1 - \int_0^t g(u) du & |\tau| = 1, \end{cases} \tag{2.1}$$

where ν_τ equals two if τ_a and τ_b are different and one if they are equal. Note that there are a discrete number of tree shapes, hence $p(\tau|t)$ is a probability

mass.

For trees with more than one species ($|\tau| > 1$) all possible lengths of the root edge are integrated over. The probability of obtaining the tree for a given λ_r is the product of the probability of the speciation event on the root lineage and the probabilities of obtaining the tree shapes descendant from the root in the remaining time. This is multiplied by two (ν_τ) if the two descendant tree shapes differ as it does not matter which descendant tree shape occurs on which lineage descendant from the root. As before the probability of obtaining a tree with a single species is simply one minus the probability of a speciation event on the root lineage occurring 'too soon'.

For any BH model, given a number of species the probability of obtaining a particular tree shape at time t is simply found by normalising over the set of all tree shapes of that size, Y_n:

$$p(\tau|t,n) = \frac{p(\tau|t)}{\sum_{\gamma \in Y_n} p(\gamma|t)}, \tag{2.2}$$

where it should be noted that τ is also in Y_n.

For Yule trees (where $g(u)$ is an exponential distribution) solutions to Equation 2.1 can be found analytically. For example it is easy to show (using induction) that the probability of obtaining a caterpillar tree (see Figure 2.1) with n leaves, $p(C_n|t)$ is:

$$p(C_n|t) = \begin{cases} 2^{n-2} e^{-t} \left(1 - e^{-t}\right)^{n-1} / (n-1)! & n > 1 \\ e^{-t} & n = 1. \end{cases} \tag{2.3}$$

Furthermore under the Yule model the tree shape probabilities ($p(\tau|n)$) are well known and independent of time (Semple and Steel, 2003):

$$p(\tau|n) = \frac{2^{n-1-s(\tau)}}{\prod_{e, c_e > 2} (c_e - 1)},$$

where $s(\tau)$ is the number of internal edges for which the two descendant trees have the same tree shape, and c_e is the number of leaf descendants of edge e. The time dependent probability for any tree shape under the Yule model

can therefore be obtained as follows:

$$\begin{aligned} p(\tau|t) &= \frac{p(\tau|t)}{p(C_n|t)} p(C_n|t) \\ &= \frac{p(\tau|n)}{p(C_n|n)} p(C_n|t) \\ &= 2^{n-1-s(\tau)} e^{-t} \left(1-e^{-t}\right)^{n-1} / \prod_{e,c_e>2} (c_e - 1). \end{aligned}$$

It is interesting to note that for a given tree size, n, the tree shape probability, $p(\tau|t)$, has a maximum at $t = \log(n)$.

For other BH models the relative tree shape probabilities for a given number of species ($p(\tau|n,t)$) may not be independent of time. To remove the time dependency contained in the relative tree shape probabilities (Equation 2.2) these probabilities must be weighted by the probability distribution of the age of a tree given that it has n leaves, which we denote by $\phi(t|n)$. Assuming a uniform prior, $p(t)$, on the age of the tree between 0 and T we can obtain the following using Bayes theorem:

$$\begin{aligned} \phi(t|n) &= \frac{p(n|t)p(t)}{p(n)} \\ &= \frac{p(n|t)p(t)}{\int_0^T p(n,u)du} \\ &= \frac{p(n|t)p(t)}{\int_0^T p(n|u)p(u)du} \\ &= \frac{p(n|t)/T}{\int_0^T p(n|u)/T du} \\ &= \frac{p(n|t)}{\int_0^T p(n|u)du} \\ &= \frac{\sum_{\gamma \in Y_n} p(\gamma|t)}{\int_0^T \sum_{\gamma \in Y_n} p(\gamma|u)du}. \end{aligned}$$

If any age is possible we can take the limit of $\phi(t|n)$ as $T \to \infty$:

$$\lim_{T\to\infty} \phi(t|n) := \phi_\infty(t|n) = \frac{\sum_{\gamma \in Y_n} p(\gamma|t)}{\int_0^\infty \sum_{\gamma \in Y_n} p(\gamma|u)du}.$$

Obviously we require the denominator to be finite, in Theorem 1 it is shown that this holds as long as the mean of the speciation probability density, $g(t)$, is finite (otherwise $p(\tau|n) = \lim_{t\to\infty} p(\tau|t,n)$ if the limit exists). Using $\phi_\infty(t|n)$ the time dependence in Equation 2.2 can be integrated out giving the time independent relative tree shape probability:

$$\begin{aligned} p(\tau|n) &= \int_0^\infty p(\tau|t,n)\phi_\infty(t|n)dt \\ &= \frac{\int_0^\infty p(\tau|t)dt}{\int_0^\infty \sum_{\gamma\in Y_n} p(\gamma|u)du}. \end{aligned}$$

This should be interpreted as the relative probability of observing a particular tree shape given that there are n species and speciation occurred according to the BH model (and the associated density, $g(t)$). One method for testing 'real' trees against such a model is to compare the distribution of tree shapes for the real trees with those predicted by the model, this is the approach taken in Blum and Francois (2006).

Theorem 1. *Consider the probability of obtaining a particular tree shape after time t, $p(\tau|t)$, under any BH evolutionary model with speciation probability density $g(t)$. If $g(t)$ has a finite mean, then the integral of $p(\tau|t)$ over all possible times is finite for any tree shape. More concisely, for any tree shape, τ, $\int_0^\infty p(\tau|t)dt$ is finite if $\int_0^\infty ug(u)du$ is finite.*

Proof. To prove Theorem 1, $p(\tau|t)$ is expressed in a new form and its integral is shown to have an upper bound of $\int_0^\infty ug(u)du$, hence if this upper bound is finite, the integral must also be finite.

Let $\phi(\tau,u)$ be the probability density that a tree shape, τ, is obtained during an evolutionary process and first occurs at time u. Let $\theta(\tau,u)$ be the probability that the tree shape exists for at least some time u. Making use of these two quantities the probability of obtaining a particular tree shape, τ, of age t can be expressed as:

$$p(\tau|t) = \int_0^t \phi(\tau,u)\theta(\tau,t-u)du. \tag{2.4}$$

This is simply the product of the probability of obtaining τ before time t and

then retaining τ until time t (no further speciation events may occur).

The probability $\theta(\tau, u)$ is complicated to derive, however for this proof an upper bound will suffice. Note that $\theta(\tau, u)$ can be interpreted as the probability that no speciation events take place in a period of length u. It will therefore be bounded by the probability that no speciation event takes place on the last lineage that speciated. Fortunately we know the time of this speciation event (at the start of the period of length u) so we have:

$$\theta(\tau, u) \leq \int_u^\infty g(v)dv.$$

Substituting in Equation 2.4 gives:

$$p(\tau|t) \leq \int_0^t \phi(\tau, u) \int_{t-u}^\infty g(v)dvdu. \tag{2.5}$$

To prove Theorem 1 we need to consider the integral of $p(\tau|t)$ over all possible times, t. Integrating both sides of Equation 2.5 and changing the order of integration we obtain the required condition:

$$\begin{aligned}
\int_0^\infty p(\tau|t)dt &\leq \int_0^\infty \int_0^t \phi(\tau, u) \int_{t-u}^\infty g(v)dvdudt \\
&= \int_0^\infty \int_u^\infty \phi(\tau, u) \int_{t-u}^\infty g(v)dvdtdu \\
&= \int_0^\infty \phi(\tau, u) \int_u^\infty \int_{t-u}^\infty g(v)dvdtdu \\
&= \int_0^\infty \phi(\tau, u)du \int_0^\infty \int_{\hat{t}}^\infty g(v)dvd\hat{t}, \qquad \text{where } \hat{t} = t - u \\
&= \int_0^\infty \int_0^v g(v)d\hat{t}dv \\
&= \int_0^\infty vg(v)dv.
\end{aligned}$$

2.2.2 Individual edge probability densities

Recently, Gernhard et al. (2006) developed a method for calculating the expected length of any edge under a Yule model. Here an alternative approach is used to give the full probability distribution of that edge length, not just

for the Yule model but for any BH model.

The probability density of the length of a particular edge, e, for a given tree shape, can be expressed recursively by integrating over the possible lengths of all other edges. To do so it is necessary to consider three possible positions of the desired edge as illustrated in Figure 2.2. Denote the probability of obtaining a tree shape, τ, at time t with a specified edge having length λ_e as $\theta(\lambda_e, \tau, t)$, making use of three possible positions of the specified edge this can be stated recursively as:

$$\theta(\lambda_e, \tau|t) = \begin{cases} \nu_\tau g(\lambda_e) p(\tau_a|t-\lambda_e) p(\tau_b|t-\lambda_e) & \text{A: root is } e \text{ and } |\tau| > 1 \\ \nu_\tau \int_0^{t-\lambda_e} g(\lambda_r)\theta(\lambda_e, \tau_a|t-\lambda_r) p(\tau_b|t-\lambda_r) d\lambda_r & \text{B: } e \text{ in } \tau_a \text{ and } |\tau| > 1 \\ \delta(t-\lambda_e) \int_{\lambda_e}^{\infty} g(u) du & \text{C: } |\tau| = 1 \end{cases} \tag{2.6}$$

In scenario A the tree (τ) contains more than one species and the desired edge is the root. The probability of obtaining the tree in this scenario is therefore simply the product of the probability of a speciation event at time λ_e on the root ($g(\lambda_e)$) and the probability of each of the daughter trees having the appropriate shape $p(\tau_a|t-\lambda_e)$ and $p(\tau_b|t-\lambda_e)$. If the two daughter tree shapes differ then there are two possible ways of obtaining the final tree shape, this introduces the factor of two (ν_τ).

In scenario B the desired edge is in one of the daughter trees which we refer to as τ_a without loss of generality. In this scenario the probability of obtaining the tree is obtained by integrating over all possible root edge lengths, λ_r. The root edge can range in length from 0 to $t-\lambda_e$ as this is the longest it can be and still 'leave' sufficient time for edge e to obtain its desired length. For a given root edge length the probability of obtaining the tree is the product of the probability of the speciation event on the root lineage ($g(\lambda_r)$), the probability that τ_a will have the appropriate shape and edge length ($\theta(\lambda_e, \tau_a|t-\lambda_r)$) and the probability that τ_b will have the appropriate shape ($p(\tau_b|t-\lambda_r)$).

In scenario C τ contains only one species which must be the desired edge. For the desired edge to have length λ_e we must have $t = \lambda_e$ hence the dirac delta function $\delta(t-\lambda_e)$. Furthermore the speciation event on the lineage

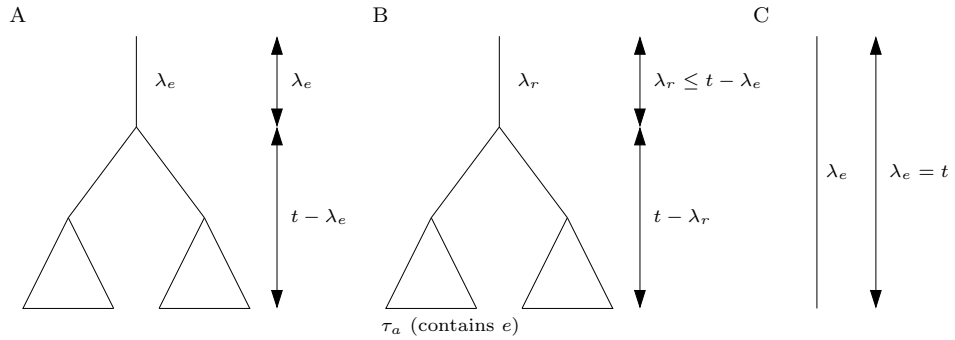


Figure 2.2: The three possible scenarios for the location of the desired edge. In Scenarios A and B the tree contains more than one species and the edge is either the root (Scenario A) or in one of the subtrees (Scenario B). In Scenario C the tree contains only one species and the pendant edge belonging to this species is the desired edge. The probability of obtaining each of these trees is considered further in the main text.

must take place after time λ_e, this gives the integral from λ_e to ∞ of the density $g(u)$ (this is simply one minus the cumulative density function of $g(u)$ at λ_e).

The probability density $\theta(\lambda_e, \tau|t)$ is not conditioned on the tree shape or the number of species. Without normalisation it should therefore be interpreted as the probability of obtaining that particular tree shape and edge length out of all possible trees of age t. In many cases it is desirable to condition $\theta(\lambda_e, \tau|t)$ on the tree shape, this is achievable by simple normalisation:

$$\theta(\lambda_e|\tau, t) = \frac{\theta(\lambda_e, \tau|t)}{\int_0^t \theta(u, \tau|t)du},$$

yielding the probability density of edge length e given the tree shape τ and age of the tree.

2.2.3 Density of the longest pendant edge

Many other interesting properties of the trees created by a BH model can be considered. In this chapter one final situation is considered where we wish to find the probability density of the longest pendant edge length for a given tree shape. The motivation for this came from a study where similarities

between different indices for biodiversity conservation were being considered (Redding et al., 2008). These indices were highly dependent on the lengths of pendant edges, consequently it became necessary to develop a good understanding of the distribution of pendant edge lengths.

A similar method to that employed in the preceding sections can be used to investigate this situation. This method is less obvious as (depending on the tree shape) there may be several edges which could be the longest pendant edge. Let τ_a and τ_b respectively be the smaller and larger daughter trees of τ. We define $\phi(\lambda_l, \tau|t)$ as the probability of obtaining a tree shape τ with a longest pendant edge of length λ_l given its age t, this can be stated recursively:

$$\phi(\lambda_l, \tau|t) = \nu_\tau \times \begin{cases} \int_0^{t-\lambda_l} g(\lambda_r) \left[\Psi(\lambda_l, t-\lambda_r, \tau_a, \tau_b) + \Psi(\lambda_l, t-\lambda_r, \tau_b, \tau_a)\right] d\lambda_r & |\tau_a| > 1 \\ g(t-\lambda_l) p(\tau_b|\lambda_l) \int_{\lambda_l}^{\infty} g(u) du & |\tau_a| = 1 \end{cases} \tag{2.7}$$

$$\Psi(\lambda_l, t, \tau_a, \tau_b) = \phi(\lambda_l, \tau_a|t) \int_0^{\lambda_l} \phi(m, \tau_b|t) dm$$

To gain some insight into Equation 2.7 we give further consideration to the two cases illustrated in Figure 2.3.

$|\tau_a| = 1$. A pendant edge is directly descendant from the root. This edge is guaranteed to be the longest pendant edge in τ. The probability, $\phi(\lambda_l, \tau|t)$, is therefore the product of the probability of the speciation event on the root lineage ($g(t-\lambda_l)$), the probability of obtaining the right tree shape for τ_b and the probability that no speciation event will take place on the pendant edge in τ_a before time t. As before the possible factor of two (ν_τ) represents the fact that it is irrelevant which set of events takes place on which of the lineages descendant from the root in τ.

$|\tau_a| > 1$. Both trees descendant from the root have more than one species. The longest pendant edge may occur in either of the trees descendant from the root, both possibilities must therefore be taken into account. $\Psi(\lambda_l, t, \tau_a, \tau_b)$ is the probability of obtaining the tree shapes τ_a and τ_b in time t with τ_a having the longest pendant edge of length λ_l. This is found by integrating over the possible longest pendant edge lengths of τ_b which can range from 0 to λ_l. Using $\Psi(\lambda_l, t, \tau_a, \tau_b)$ the probability of obtaining the tree τ with longest

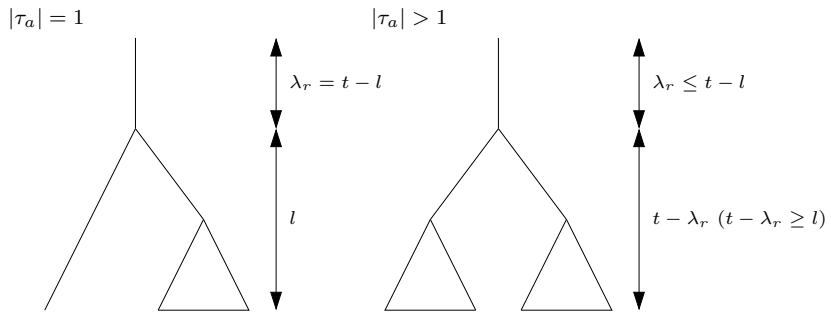


Figure 2.3: The two possible scenarios considered for the location of the longest pendant edge. Either a pendant edge is adjacent to the root, in which case it must be the longest, or it may be located in either of the trees descendant from the root.

pendant edge λ_l is expressed as the integral over all possible root lengths, λ_r, of the product of the probability of obtaining a speciation event at time r and a longest pendant λ_l from one of the trees descendant from the root.

2.2.4 Analytic and numerical considerations

For the Yule model the equations for $p(\tau|t)$, $p(\tau|n)$, $\theta(\lambda_e, \tau|t)$ and $\phi(\lambda_l, \tau|t)$ possess analytic solutions. This is easily proven using induction on a case by case basis. The solutions for the Yule model are analytic because at each stage in the recursive approach a sum of exponential terms with linear exponents are produced, these expressions are easily combined and integrated to obtain an expression with the same form.

A symbolic algebra package was used to solve the recursive equations we have presented for the Yule model. Unfortunately not all BH models possess analytic solutions to the recursive equations. For instance if $g(t)$ is a distribution that does not possess an analytic cumulative density function then even the simplest case for $p(\tau, t)$ with $|\tau| = 1$ will not possess an analytic solution. To obtain solutions for the methods presented in this chapter it may therefore be necessary to resort to numerical methods; this is further complicated by the nested nature of the integrals.

2.3 Approach 2: Using rank functions for Yule trees

The second approach we present utilises rank functions for Yule models as introduced in Gernhard et al. (2006); Gernhard (2006). No work to date has been done on rank functions of BH models in general, hence at present this approach is strictly for Yule models. Using rank functions we derive a closed form equation for the density of an edge length in a tree which evolved under the Yule model, $\theta(\lambda_e|\tau, t)$.

For the concept of rank functions, we need to consider a tree shape with leaf labels, a so called phylogenetic tree $\mathcal{T}$. In the following, if a tree shape is given, we label the leaves in an arbitrary way to obtain a phylogenetic tree. Let $\mathring{V}$ be the set of vertices in $\mathcal{T}$ of degree > 1. So the set $\mathring{V}$ consists of all vertices in $\mathcal{T}$ except of leaves and the root of the tree. A rank function (Semple and Steel, 2003) on a phylogenetic tree is a bijection from $\mathring{V}$ to $\{1, 2, \ldots, |\mathring{V}|\}$ with the property that the ranks are increasing on any path from the root to a leaf. We call a phylogenetic tree with a rank function a ranked phylogenetic tree.

The Yule model induces a uniform distribution on the ranked phylogenetic trees on n species (Aldous, 2001). In Gernhard et al. (2006); Gernhard (2006), polynomial time algorithms for calculating the probability of the rank of a vertex are provided for the uniform distribution on ranked phylogenetic trees. In the following we will explain the idea of the algorithms and adjust them to the application in this chapter.

2.3.1 Calculating the rank distribution

Let r be a rank function on the phylogenetic tree $\mathcal{T}$. Define $p_u := (\mathbb{P}[r(u) = i])_{i=1,\ldots,n-1}$. In Gernhard et al. (2006), a formula for calculating p_u is given: Label the vertices on the path from the vertex u to the most recent common ancestor $mrca$ with $u = x_1, x_2, \ldots, x_m = mrca$, see Fig. 2.4. Define λ_j as the number of leaves below x_j minus 1. With that notation, we get from Gernhard et al. (2006) that

$$p_u = \frac{M_{m-1} M_{m-2} \ldots M_1 e_1}{|M_{m-1} M_{m-2} \ldots M_1 e_1|_1} \tag{2.8}$$

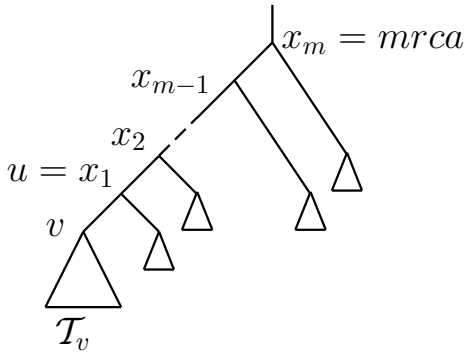


Figure 2.4: Labeling of the tree for calculating the probability for the rank of a vertex.

where $|\cdot|_1$ is the 1-norm, $e_1 = (1, 0, 0, \ldots, 0)^T$ and the matrix M_j is defined as follows,

$$(M_k)_{i,j} = \begin{cases} 0 & \text{if } j < i - 1 - (\lambda_{k+1} - \lambda_k), \\ 0 & \text{if } j > i - 1, \\ \binom{\lambda_{k+1}-i}{\lambda_{k+1}-\lambda_k-i+j+1}\binom{i-2}{i-j-1} & \text{otherwise.} \end{cases}$$

The algorithm RankProb in Gernhard et al. (2006) calculates p_u according to Equation 2.8.

For an edge $e = (u, v)$ in $\mathcal{T}$, we want to obtain the probability $p_{u,v}(i, j) := \mathbb{P}[r(u) = i, r(v) = j]$. First, let e be an interior edge. In Gernhard et al. (2006), we calculate $p_{u,v}(i, j), 1 \leq i < j \leq n - 1$ by running RankProb on different subtrees of $\mathcal{T}$. In the following, we give an expression to calculate $p_{u,v}(i, j)$ directly from $p_u(i)$ which makes the calculations faster. Let $\mathcal{T}_v$ be the smallest subtree induced by the leave descendants of v, see Fig. 2.4. The subtree $\mathcal{T}_v$ has n_v leaves. Let $r(\mathcal{T})$ be the set of rank functions on $\mathcal{T}$.

The number of rank functions where $r(u) = i$ is $p_u(i) \cdot |r(\mathcal{T})|$. Assume we fix the first i interior nodes, with u being the ith node. There are $\binom{n-1-i}{n_v-1}$ possibilities to shuffle the interior vertices in $\mathcal{T}_v$ with the remaining interior vertices. Only $\binom{n-1-j}{n_v-2}$ of those shuffles assign rank j to vertex v. Overall, we

therefore get for the number of rank functions with $r(u) = i$ and $r(v) = j$:

$$p_u(i) \cdot |r(\mathcal{T})| \frac{\binom{n-1-j}{n_v-2}}{\binom{n-1-i}{n_v-1}}$$

For the probability $p_{u,v}(i,j)$, we have to divide the previous equation by the number of rank functions. Therefore

$$p_{u,v}(i,j) = p_u(i) \cdot \frac{\binom{n-1-j}{n_v-2}}{\binom{n-1-i}{n_v-1}}$$

This is equivalent to

$$p_{u,v}(i,j) = \begin{cases} p_u(i) \frac{n_v-1}{n-n_v-i+1} \prod_{k=1}^{n_v-2} \frac{n-j-k}{n-i-k}, & \text{if } n-j+1 \geq n_v, 1 \leq i < j < n; \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

We will extend the distribution $p_{u,v}$ for leaves. Since the leaves are after the $(n-1)$st speciation event, we can assume that all leaves have rank n. So for pendant edges, we have

$$p_{u,v}(i,n) = \begin{cases} p_u(i), & \text{if } v \text{ is a leaf;} \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

Further, we will define $p_r, p_{r,v}$ for the root r. The root is always the very first vertex, the most recent common ancestor ($mrca$) is its descendant. Therefore, we define,

$$p_r(0) = 1, \qquad p_{r,v}(0,1) = \begin{cases} 1, & \text{if } v \text{ is the } mrca; \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

2.3.2 Calculating densities under the Yule model

In the following we want to calculate the edge length density for an edge $e = (u, v)$ in a tree τ. Let Λ_e be the random variable 'edge length of e with density function θ. Let Λ_u be the random variable 'time of the speciation event u' with density function f_{Λ_u}. In a tree with n species, let $\Lambda_{i,j}$ be the random variable 'time between the $i-$th and the $j-$th speciation event';

with density function $f_{\Lambda_{i,j}}$. Let Λ_i be the random variable 'time of the $i-$th speciation event'; with density function f_{Λ_i}. With $i = 0$, we denote the root of the tree. Time is measured between today and the speciation event.

Under the Yule model, the waiting time between the $(k-1)$st speciation event and the kth speciation event, X_k, is exponentially distributed with rate k. We have

$$\Lambda_{i,j} = \sum_{k=i+1}^{j} X_k. \tag{2.12}$$

The present is between the $(n-1)$st speciation event and the nth speciation event. However, it has been shown in Hartmann et al. (a) that the time between the $(n-1)$st speciation event and the present has the exponential (rate n) distribution, this is X_n. So the present can be considered as the nth speciation event when not conditioning on the age of the tree, t. Later we will see that the same holds with conditioning on t.

We will derive the density and expectation of the random variables $\Lambda_i, \Lambda_{i,j}, \Lambda_u, \Lambda_e$. Recall that u is some interior vertex and e any edge. For the density, we have,

$$\begin{aligned} f_{\Lambda_u}(\lambda_u|\tau) &= \sum_{i=0}^{n-1} f_{\Lambda_i}(\lambda_u|n)p_u(i); \\ \theta(\lambda_e|\tau) &= \sum_{i=0}^{n-1}\sum_{j=i+1}^{n} f_{\Lambda_{i,j}}(\lambda_e|n)p_{u,v}(i,j). \end{aligned}$$

Note that for interior edges, $p_{u,v}(i,n) = 0$, for pendant edges, $p_{u,v}(i,j) = 0$ for $j < n$ and for the root edge, $p_{u,v}(0,1) = 1$.

For the expectation, we obtain from $\mathbb{E}[\Lambda_i|n]$,

$$\begin{aligned} \mathbb{E}[\Lambda_u|\tau] &= \sum_{i=0}^{n-1} \mathbb{E}[\Lambda_i|n]p_u(i); \\ \mathbb{E}[\Lambda_{i,j}|n] &= \mathbb{E}[\Lambda_i|n] - \mathbb{E}[\Lambda_j|n], \qquad 0 \le i < j \le n; \\ \mathbb{E}[\Lambda_e|\tau] &= \mathbb{E}[\Lambda_u|\tau] - \mathbb{E}[\Lambda_v|\tau] = \sum_{i=0}^{n} \mathbb{E}[\Lambda_i|n](p_u(i) - p_v(i)); \end{aligned}$$

with Λ_n being the present (which is, as explained above, equivalent to Λ_n

being the time of the n-th speciation event). Hence, $\Lambda_n = 0$ and therefore $\mathbb{E}[\Lambda_n|n] = 0$.

We can calculate the probabilities $p_u(i), p_{u,v}(i,j)$ as described in section 2.3.1, so it is left to calculate $f_{\Lambda_{i,j}}(\lambda_{i,j}|n), f_{\Lambda_i}(\lambda_i|n), \mathbb{E}[\Lambda_i|n]$. The three values will be computed both conditioning and not conditioning on the age of the tree.

Unknown age of the tree

If we do not condition on the age of the tree, t, but assume a uniform prior for the age of the tree, the distribution of $\Lambda_i, \Lambda_{i,j}$ was calculated by Gernhard (2008) as follows,

$$f_{\Lambda_i}(\lambda_i|n) = (i+1)\binom{n}{i+1}e^{-n\lambda_i}(e^{\lambda_i}-1)^{n-i-1}, \quad 0 \le i \le n-1; \tag{2.13}$$

$$\mathbb{E}[\Lambda_i|n] = \sum_{k=i+1}^{n}\frac{1}{k}, \qquad 0 \le i \le n-1;$$

$$\begin{aligned} f_{\Lambda_{i,j}}(\lambda_{i,j}) &= f_{\Lambda_{i,j}}(\lambda_{i,j}|n) \\ &= (i+1)\binom{j}{i+1}e^{-j\lambda_{i,j}}(e^{\lambda_{i,j}}-1)^{j-i-1}, \quad 0 \le i < j \le n. \end{aligned} \tag{2.14}$$

Note that $\Lambda_{i,j}$ and $\Lambda_{k,l}, i < j \le k < l$ are independent since the X_k are independent.

Known age of the tree

Next we state the density and expectation of Λ_i and the density of $\Lambda_{i,j}$ conditioned on the age of the tree. The proofs of these three theorems are provided in Gernhard, Hartmann and Steel.

Theorem 2. *Each random variable Λ_i $(1 \le i \le n-1)$ has the density function:*

$$f_{\Lambda_i}(\lambda_i|n,t) = i\binom{n-1}{i}(1-e^{-t})^{1-n}e^{-i\lambda_i}(1-e^{-\lambda_i})^{n-i-1}(1-e^{-(t-\lambda_i)})^{i-1}.$$

For $i = 0$, we have $f_{\Lambda_0}(\lambda_0|n,t) = \delta(\lambda_0 - t)$.

Theorem 3. *The expectation of Λ_i (for $1 \leq i \leq n-1$) is:*

$$\mathbb{E}[\Lambda_i|n,t] = \sum_{k_1=0}^{n-i-1} \sum_{k_2=0}^{i-1} B_{k_1,k_2}(1-e^{-t})^{1-n}(e^{-k_2 t} - ((i+k_1-k_2)t+1)e^{-(i+k_1)t})$$

with $B_{k_1,k_2} := i\binom{n-1}{i}\binom{n-i-1}{k_1}\binom{i-1}{k_2}(-1)^{k_1+k_2}(i+k_1-k_2)^{-2}$.
For $i = 0$, we have $\mathbb{E}[\Lambda_0|n,t] = t$.

Theorem 4. *Each random variable $\Lambda_{i,j}$ ($1 \leq i < j \leq n-1$) has the density function:*

$$\begin{aligned} f_{\Lambda_{i,j}}(\lambda_{i,j}|n,t) &= \sum_{k_1=0}^{i-1} \sum_{k_2=0}^{n-j-1} B_{k_1,k_2} e^{(n-j)\lambda_{i,j}} \frac{(e^{\lambda_{i,j}}-1)^{j-i-1}}{(e^t-1)^{n-1}} \times \\ &\quad (e^{(n-i+k_1)(t-\lambda_{i,j})} - e^{k_2(t-\lambda_{i,j})}) \end{aligned}$$

with $B_{k_1,k_2} = i(i+1)\binom{j}{i+1}\binom{n-1}{j}\binom{i-1}{k_1}\binom{n-j-1}{k_2}\frac{(-1)^{n+i-j-k_1-k_2}}{n-i+k_1-k_2}$.
For $\Lambda_{i,n}, i < n$, we have $f_{\Lambda_{i,n}}(\lambda_{i,n}|n,t) = f_{\Lambda_i}(\lambda_{i,n}|n,t)$, i.e. today can be interpreted as the n-th speciation event.
For $\Lambda_{0,j}$ we have $f_{\Lambda_{0,j}}(\lambda_{0,j}|n,t) = f_{\Lambda_1}(t - \lambda_{0,j}|n,t)$.

2.4 Concluding comments

In this chapter we have studied the class of Bellman Harris (BH) evolutionary models, a class that includes the widely used Yule model. A method for calculating various probability distributions of tree shapes and edge lengths of trees produced under BH models has been presented. For Yule models analytic solutions exist for the proposed method, however for other BH models it may be necessary to resort to numerical methods.

A second method for calculating edge densities using rank functions has also been presented. This method only applies to the Yule model, however this limited scope makes this method conceptually easier to work with and implement. Reassuringly for the Yule model our methods (which are conceptually independent) give identical results.

Obtaining analytic solutions for properties of BH models in general or Yule models in particular can be complicated. However results often exist – particularly for Yule models – and it is worthwhile pursuing these. The approaches presented here are adaptable to a wide range of questions and scenarios. If it is necessary or desirable to simulate phylogenetic trees instead we caution that this should be done with some care using approaches such as those discussed in chapter 3 and Hartmann et al. (a).

Chapter III

Sampling trees from evolutionary models

In the previous chapter we considered mathematical characteristics of Bellman-Harris models and the Yule model. For other models and applications analytical solutions may be difficult to obtain or may not exist. To overcome this issue a sample of trees can be obtained from the evolutionary model and used to investigate the problem. In this chapter we show why a commonly used sampling approach should only be applied to certain models and provide an alternative approach that can be applied to most other models. We explore the bias produced by inappropriate sampling methods and identify situations in which this bias is particularly pronounced. The algorithms presented here are available in the Perl Bio::Phylo package and as a stand-alone program TreeSample. This chapter has been submitted (in a slightly different form) as Hartmann et al. (a).

For most models sampling trees appears to be a relatively easy exercise. If the aim is to produce trees of a given age this is indeed true. However in many circumstances it is preferable to sample trees with a given number of species. There are numerous ways to produce trees with a given number of species from an evolutionary model, however many seemingly intuitive approaches sample trees from unexpected and unrealistic distributions. This introduces some potential pitfalls, a problem that is exacerbated by the fact that there is no easy method for testing whether the sampling approach is correct.

Some simple approaches for sampling trees with a given number of species are in common usage. In this chapter we show that these approaches are appropriate for the widely used Yule and Coalescent models but there are some fundamental problems applying these approaches to other evolutionary models. We provide alternative sampling approaches that are theoretically

sound and easy to apply.

We investigate the importance of using our correct sampling approach over established methods. This is achieved by comparing samples of trees produced by the different sampling approaches for any given model. Existing sampling approaches introduce a strong bias in the age of a tree and a less pronounced bias in the relative timing of the speciation events. For the considered models, existing approaches introduce a negligible bias in the tree shape distribution and for incomplete taxon sampling. We identify attributes of other models that will result in existing sampling approaches producing more biased samples.

The methods we present are not the fastest or most sophisticated, however in our opinion they are the easiest to implement and applicable to the broadest possible range of models. Most of our algorithms are implemented in the Perl Bio::Phylo package, where they can easily be applied to any suitable evolutionary model. For those users unfamiliar with Perl we have also made them available using a stand-alone GUI TreeSample. These tools are freely available from our website (Hartmann, 2007). Lastly we note that although we present our work in the context of evolutionary models of species diversification, our methods can be applied to other scenarios where birth-death processes are modeled, for example gene trees (Oakley et al., 2006; Karev et al., 2003; Hahn et al., 2005).

3.1 Sampling methods

Throughout this chapter we assume that we want to produce a sample from the tree probability distribution induced by an evolutionary model. The first problem is that this tree probability distribution is ill defined for most evolutionary models. Under most models trees evolve perpetually and trees of all ages are possible, hence the expected age of the tree (the time between the root and the tips) is infinite. To obtain a probability distribution it is therefore necessary to condition on some aspect of the tree; the number of species or the age of the tree are arguably the two most common and useful choices.

Conditioning on the age of a tree is appropriate if we wish to compare a

model with trees of known age or want to test methods on simulated trees of a given age. It is relatively easy to sample trees of a given age from an evolutionary model. The tree is simply evolved according to the model until it has reached the desired age. This process is repeated until a sufficient number of trees have been sampled.

Conditioning on the number of species, n, in a tree may be of more interest for real applications. The age of a constructed tree may only be known with limited accuracy, however the number of species in the (constructed) tree is fixed. Consequently it may be more appropriate to use samples from an evolutionary model with a fixed number of species (we also consider incomplete taxon sampling). Sampling from the tree distribution conditional on the number of species, $p(\mathcal{T}|n)$, is the basis of this chapter.

Throughout this chapter we assume a uniform prior on the age of the tree as done in Aldous and Popovic (2005); Popovic (2004); Gernhard, Hartmann and Steel. Consider a large number of simulation runs that begin at a uniformly distributed time before the present. Trees obtained by selecting only those simulations that have n species at the present are a sample from $p(\mathcal{T}|n)$. This is a convenient way of interpreting the distribution but is not a practical sampling approach as the simulation starting time is taken from an ill defined distribution (between an infinite time in the past and the present). A given model (and its parameters) will induce a distribution on the age of the tree given its size. All our knowledge about the age of a tree is encapsulated in the model and the chosen parameter values; the uniform prior on the tree age represents the fact that we have no further knowledge about the tree age outside of these parameters.

3.1.1 Current approaches

One simple sampling approach (which we refer to as *SSA*) for sampling trees with n species has seen wide usage. With this approach a tree is evolved under the model until it has $n + 1$ species and the last speciation event is disregarded. This approach produces trees conditional on the next speciation event occurring immediately after the end of the tree, which as we show here is generally not the same distribution as $p(\mathcal{T}|n)$. It is difficult to justify this

approach as it produces a sample of trees equivalent to what we would expect if all 'real' trees were observed immediately prior to a speciation event.

PhyloGen (Rambaut, 2002) is a freely available tree sampler that has been used in a number of studies (eg. Hohl and Ragan (2007); Shaw et al. (2003); Venditti et al. (2006); Weir (2006)). It permits users to sample trees from constant rate birth-death processes and episodic speciation models. These trees are conditioned on the age of the tree or the number of species, n. Conditioning on n in *PhyloGen* simply terminates a tree after it first reaches n species. Trees sampled with *PhyloGen* are younger than expected for our interpretation of $p(\mathcal{T}|n)$ and the pendant edges are shorter than expected – in fact the species produced by the last speciation event have zero length edges. If the last speciation event is removed (creating a tree with $n-1$ species) sampling trees with *PhyloGen* is equivalent to *SSA* with $n-1$ species. Due to this similarity throughout the remainder of this chapter we only consider *SSA*.

There are three main possible problems with *SSA* and *PhyloGen*:

Problem 1. As has already been noted the pendant edge lengths produced by *SSA* and *PhyloGen* have what appears to be extreme values. With *PhyloGen* the pendant edges are as short as possible and with *SSA* they seem too long (this will be discussed in more detail later).

Problem 2. *SSA* and *PhyloGen* stop evolving the tree during (or just after) the first period of time where the tree has n species. For models with extinction the number of species will fluctuate up and down so there may be many periods during which the tree has n leaves. For such models *SSA* and *PhyloGen* will result in younger trees than expected.

Problem 3. A final concern with *SSA* and *PhyloGen* is that each model simulation run makes the same contribution to the final sample – one single tree. However, from our interpretation of $p(\mathcal{T}|n)$ the probability of observing a given simulation depends on the duration for which the simulated tree had n species – for example, if this duration is short it is unlikely that the simulated tree will be observed whilst it has n species.

3.1.2 Pure birth memoryless models

We begin by considering pure-birth memoryless models – models that do not explicitly include extinction (pure birth) and where future evolution depends only on the number of extant species (memoryless). This class of models is of particular interest as an approach similar to *SSA* can be used to correctly sample phylogenetic trees from them. Furthermore this class of models includes the most widely used speciation model – the Yule model (Yule, 1924; Harding, 1971) – and the most widely used null model in population genetics – the Coalescent (Kingman, 1982a,b,c).

Under the Yule model each species has the same probability of speciating per unit time and this speciation rate is constant over time. Consequently the time between speciation events is exponentially distributed with parameter $m\beta$, where m is the number of species that are extant and β is the intrinsic rate of speciation. The Coalescent is derived from population genetics principles but is essentially the same as the Yule model with one exception – the time between coalescent events is exponentially distributed with parameter $\binom{m}{2}$ (in the following we will use 'speciation' for both speciation and coalescent events).

In this section we show that although *SSA* is generally inappropriate for Pure birth memoryless models it is actually a correct approach for the Yule model and the Coalescent. As these models are pure birth models there will only be one period during which n species exist, so *Problem 2* does not apply. This leaves *Problems 1* and *3* which we will show cancel each other out under the Yule model and the Coalescent. We speculate that the suitability of *SSA* for sampling from the most widely used null models has led to its application to other models for which it is unsuitable.

An important aspect of memoryless models is that the evolution after the speciation event that created the nth species (s_n) is completely independent of the evolution that occurred up to that point. Consequently it is possible to simulate trees from these models in two separate stages. Firstly using the model a tree is simulated to the speciation event that created the nth species (denoted by s_n; see Figure 3.1). A length λ is then added onto the pendant edges to produce the final tree. Due to the independence of these

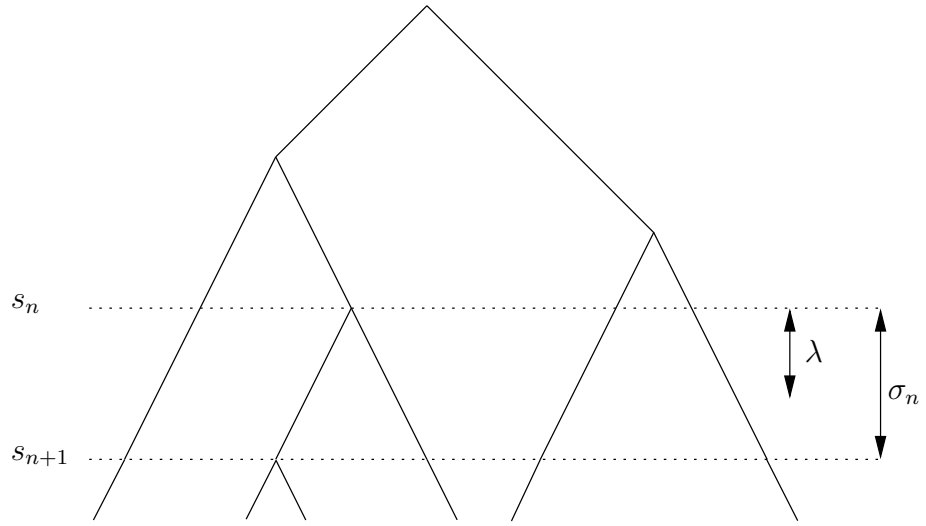


Figure 3.1: Some of the notation used throughout this chapter is illustrated in this figure where $n = 5$. τ is the simulated tree until the point in time when the tree first has greater than n species. This point in time is the speciation event creating the $(n+1)$th species – s_{n+1}. The duration for which a simulated tree has n species is denoted by σ_n, this is the time between the creation of the nth species (s_n) and the $(n+1)$th species (s_{n+1}). The time for which an observed tree has n species is necessarily less than σ_n and is denoted by λ.

two processes, *Problems 1* and *3* do not effect the simulation to s_n and are addressed entirely by an appropriate choice of λ. This raises the question from what probability density, $h(\lambda)$, the additional time λ should be sampled.

We begin by noting that any pure-birth memoryless model can be uniquely defined by the probability densities of the intervals between speciation events. We denote the time between the speciation event that created the nth and the $n+1$th species by σ_n (the time between s_n and s_{n+1}) and its probability density by $g_n(\sigma_n)$.

Note that *SSA* makes the assumption that

$$h(\lambda) = g(\lambda).$$

this effectively produces a tree with n species conditional on the next speciation event occurring immediately – clearly not what was intended.

A seemingly better (but still generally incorrect) approach would be to simulate the tree until s_{n+1} and randomly terminate the tree between s_n and s_{n+1} (since all trees between these two events should be equally likely). This addresses *Problem 1* and gives us:

$$\begin{aligned} h(\lambda) &= \int_{\sigma_n=\lambda}^{\sigma_n=\infty} h(\lambda|\sigma_n) g_n(\sigma_n) d\sigma_n \\ &= \int_{\sigma_n=\lambda}^{\sigma_n=\infty} \frac{g_n(\sigma_n)}{\sigma_n} d\sigma_n \end{aligned}$$

However this is does not take into account the variable contribution to the $p(\mathcal{T}|n)$ that different values of σ_n should make (*Problem 3.*).

From the definition of $p(\mathcal{T}|n)$ the contribution from a simulated tree with a given σ_n should be proportional to σ_n, therefore the correct distribution from which to sample λ is:

$$\begin{aligned} h(\lambda) &\propto \int_{\sigma_n=\lambda}^{\sigma_n=\infty} \sigma_n h(\lambda|\sigma_n) g_n(\sigma_n) d\sigma_n \\ h(\lambda) &\propto \int_{\sigma_n=\lambda}^{\sigma_n=\infty} g_n(\sigma_n) d\sigma_n \end{aligned} \tag{3.1}$$

Thus the following will produce correct samples from $p(\mathcal{T}|n)$ for any pure-birth memoryless model:

Pure-birth memoryless sampling approach (*PBMSA*)

1. Simulate a tree terminating at s_n
2. Add a distance, λ, to the pendant edges using the correct $h(\lambda)$ from Equation 3.1
3. Repeat from step 1 until all samples are obtained

For *SSA* to be appropriate we require $h(\lambda) = g_n(\lambda)$. Inspection of Equation 3.1 reveals that this requirement is met if $g_n(\sigma_n)$ is an exponential distribution. Furthermore as the model is memoryless the parameter may depend only on the number of species that are extant. These conditions are clearly satisfied by the Yule model, the Coalescent and the related Moran (Moran, 1958) and Hey models (Hey, 1992).

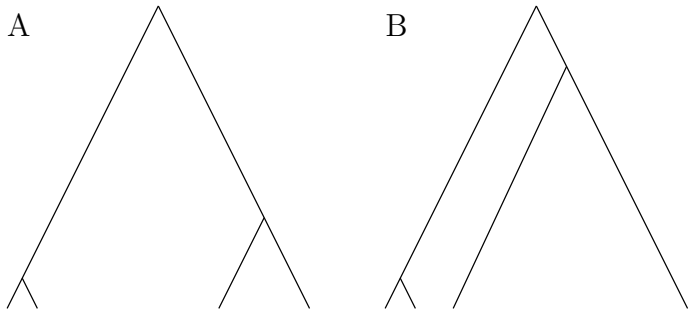


Figure 3.2: Consider an evolutionary model where 'young' species have a higher chance of speciating. Under this model the tree in A is expected to have four species for a shorter duration than the tree in B. The tree in B should therefore make a greater contribution to our sample if we want to sample trees from the model conditional on them having four leaves.

PBMSA is appropriate for any model where the time between speciation events depends only on the number of extant species, however the Yule model and the Coalescent are the only widely used models that fit this category. *PBMSA* is inappropriate for models with explicit extinction events and models with a memory. Explicit extinction events will result in a simulated tree that may have n species for several intervals – *PBMSA* would only sample from the first of these intervals resulting in a tree that is younger than expected.

Many models feature a memory, this may be in the form of hereditary speciation rates (e.g. Heard (1996)) or a dependence of speciation rates on the absolute age of a tree or a species (e.g. Chan and Moore (1999)). *PBMSA* cannot sample from such models as the evolution before and after s_n is not independent and different simulations to s_n should make different contributions to the final sample. Consider a model where young species are much more likely to speciate than their older counterparts. Figure 3.2 shows two simulated trees to s_n where $n = 4$. In A there are four young species, in B there are only two young species (those produced at s_n). Consequently tree A is expected to have n species for a shorter time than B and by the definition of $p(\mathcal{T}|n)$ should give a greater contribution to that density. Consequently it is necessary to take different numbers of samples from each of the evolutionary histories and *PBMSA* cannot be used.

3.1.3 A general sampling approach

We now introduce a general sampling method that works for a broad class of models that can include both speciation and extinction events. Our sampling approach simulates a tree, τ, until it is highly unlikely that the tree will return to n species. This will occur either when all species are extinct or when there has been sufficient speciation such that the number of extinctions required to return to n species are highly improbable.

The only restriction on the class of models from which our algorithm can sample is that we must be able to guarantee that each simulation 'run' will eventually terminate. The efficiency of the algorithm depends on the time that is required until a simulation terminates. An example of a model to which this algorithm can not be applied is one where the number of species perpetually fluctuates over a range including n.

Determining how unlikely a tree is to return to n species depends on the model. Throughout the remainder of this section we assume that we can determine a critical number of species, n^*, from which it is unlikely that extinctions will bring the number of species back to n. A simulation therefore ends when the number of species reaches 0 or n^*. n^* can be obtained via simulations.

For some models the termination condition may be much more complicated, consider a model with evolving speciation and extinction rates – an appropriate termination condition will depend both on the number of species and on the speciation and extinction rates.

A simulation run will have k periods during which n species were extant, we denote the length of each of these periods by $\phi_i, i = 1, \ldots, k$. As previously discussed, the probability of observing a simulated tree whilst it has n species is directly proportional to the duration for which n species existed: $\Phi = \sum_{i=1}^{k} \phi_i$. This will vary between simulations so each simulation should make a different contribution to the final sample – a simulated tree where n species existed for a short period of time should make a lower contribution to the sample than a simulated tree where n species existed for a longer period.

The question remains how to decide on the number of samples to take from a given simulated tree, this should be proportional to Φ. To take

this into account we introduce a sampling rate, r, such that we will take $r\Phi$ sampled trees from a given simulated tree. As we can only take whole samples of trees, for each simulated tree $r\Phi$ will be randomly rounded: If $r\Phi$ is between integers k and $k+1$, it is rounded down with probability $r\Phi - k$ and up with probability $1-(r\Phi-k)$. This ensures that the randomly rounded $r\Phi$ has an expected value of $r\Phi$.

If the sampling rate is too low many simulations will be required for each sampled tree and the process will be very inefficient. If it is too high many sampled trees may be derived from a single simulated tree and these sampled trees will have a higher degree of correlation than expected for random samples. Ideally r should be determined experimentally (by simulations) such that it is as high as possible whilst ensuring that few simulated trees produce more than a single sample. Like n^*, an appropriate value for r can be obtained from simulations.

Lastly we introduce $S_i(\tau)$ as the set of trees that can be obtained by truncating a simulated tree during the ith interval during which it had n species. Combining these elements we have the following sampling approach:

General sampling approach (*GSA*)

1. Determine a suitable sampling rate, r
2. Simulate a tree, τ, until n^* species or extinction is reached
3. Find the expected number of trees to sample from τ: $r\Phi = \sum_{i=1}^{k} r\phi_i$
4. Randomly round $r\Phi$
5. For each sample required:
   (a) Randomly choose an interval, i, according to the weights ϕ_i
   (b) Sample a tree uniformly at random from $S_i(\tau)$
6. Repeat from step 2 until the required number of samples have been obtained

3.1.4 Extension of GSA *to incomplete taxon sampling*

Most n species trees based on real data will be a subsample of the m species contained in the true underlying tree, such that $m-n$ species are missing. This problem is referred to as incomplete taxon sampling (see e.g.

Zwickl and Hillis (2002)) and may be due to several reasons including inability to sample the species or a species being 'undiscovered'. If the number of species that are missing in a tree is substantial, incomplete taxon sampling should be included explicitly. A common approach is to sample trees with m species and randomly remove $m - n$ species, thus producing an n species tree as desired. For example, if only 75% of species are being sampled and we wish to sample a tree with 30 species, we would generate a tree with 40 species and remove 10 species uniformly at random. The problem with this approach is that we will generally only have an estimate of the number of missing species (25% in our example), hence we should consider a range of possible missing numbers of species. For instance in the previous example the true tree may have somewhere between, say, 35 and 50 species.

Here we extend *GSA* to explicitly take into account incomplete taxon sampling. This extension of *GSA* requires either an estimate of the probability, s, of any given species being sampled or alternatively the probability distribution of the size of the true tree, m, given the number of sampled species, n. Without one of these quantities our method cannot be applied and indeed, it is difficult to see how to proceed otherwise. Our method also assumes that sampled species are uniformly at random distributed through the tree. It is relatively straightforward to relax this last assumption, although we do not present any details here. One instance where this would be necessary is if the probability of sampling any two species is positively correlated to their proximity in the phylogenetic tree (as might be the case if whole clades are likely to be missed, or thoroughly sampled).

Given the sampling probability s, for a given real tree size, m, the number of sampled species, n, will be distributed according to a binomial distribution:

$$p(n|m) = \binom{m}{n} s^n (1-s)^{m-n} \tag{3.2}$$

However the number of sampled species, n, is the size of the final tree and is what we wish to condition on, thus Bayes' Law gives us:

$$p(m|n) \propto p(n|m)p(m) \tag{3.3}$$

where $p(m)$ is the probability of a tree having m leaves and $p(n|m)$ is the

probability of sampling n of those leaves. For $m \geq n$ it is always possible to to obtain n leaves from a tree with m leaves, however the probability of this occurring decreases with m, such that $p(n|m)$ becomes small enough to make $p(m|n)$ negligible. This permits us to restrict the range of m that must be examined to $n \leq m \leq m^*$ where m^* is a limit that needs to be established. If we assume that $p(m)$ does not increase with m, an appropriate condition to solve for m^* is:

$$p(n|m^*) \leq \sum_{m=n}^{m^*-1} \frac{p(n|m)}{N}, \tag{3.4}$$

where N is the number of trees which are being sampled. This condition ensures that the first value of m being excluded is expected to contribute less than one tree to the final sample. If $p(m)$ increases with m extra analysis will be required to find an appropriate m^* (eg. using simulation studies).

Given a particular simulated tree we have $p(m) \propto \Phi_m$ (the duration for which a simulated tree had m species), hence substitution in Equation 3.3 gives:

$$p(m|n) \propto \Phi_m \binom{m}{n} s^n (1-s)^{m-n}, \tag{3.5}$$

which is readily normalised to give $p(m|n)$. The expected contribution to the sample from a given simulated tree consists of the expected contribution for each value of m:

$$r \sum_{m=n}^{m^*} \Phi_m p(m|n). \tag{3.6}$$

When a tree is simulated, the expected contribution to the sample is found and a sample of the corresponding size is taken. This process is repeated until the sample has the desired size.

GSA with incomplete taxon sampling

1. Find m^* analytically or by simulation / investigation (eg. Equation 3.4)
2. Simulate a tree, τ until m^* species are reached or all species become extinct

3. Calculate $p(m|n)$ for all m for this simulated tree (Equation 3.5)
4. Find the expected number of samples to take from τ (Equation 3.6)
5. Randomly round the expected number of samples
6. For each sample:
   (a) Randomly choose the original tree size, $\hat{m}$ according to $p(m|n)$
   (b) Uniformly at random choose a time when τ had $\hat{m}$ species
   (c) Randomly delete $\hat{m} - n$ species
7. Repeat from step 2 until all samples have been obtained

3.2 Efficient sampling from the constant rate birth-death model

In this section we present an efficient algorithm for sampling trees with n species from the constant rate birth-death model. The constant rate birth-death model is a popular null model for detecting variation in diversification rates (e.g. Pybus and Harvey (2000), Mooers and Heard (1997) and Chan and Moore (2002)). It is an extension of the Yule model where all species have a constant rate of speciation, β, and a constant rate of extinction, μ, with the constraint that $\beta \geq \mu$.

The method we propose relies on representing a binary tree as a point process, this is illustrated in Figure 3.3. Generally, a binary tree with n extant species can be described by $n-1$ points in the following way. On a horizontal axis, locate the leaves (species) at $1, 2, \ldots n$. The $n-1$ speciation times are represented by $n-1$ points with (x, y) co-ordinates $(j + 1/2, s_j)$, $j = 1, 2, \ldots, n; s_j > 0$. The tree is obtained by an iterative procedure. At each step of the iteration the most recent speciation event is connected with the two neighboring leaves. This speciation event is regarded as a new leaf and replaces the two neighboring leaves. This is repeated until all speciation points are connected.

In Gernhard (2008), it is shown that the times s_i of the speciation events in a constructed tree under the constant rate birth-death model are independent and identically distributed. For $\beta > \mu$, we have the distribution function:

$$F(s|t, \beta, \mu, n) = \frac{1 - e^{-(\beta-\mu)s}}{\beta - \mu e^{-(\beta-\mu)s}} \frac{\beta - \mu e^{-(\beta-\mu)t}}{1 - e^{-(\beta-\mu)t}}$$

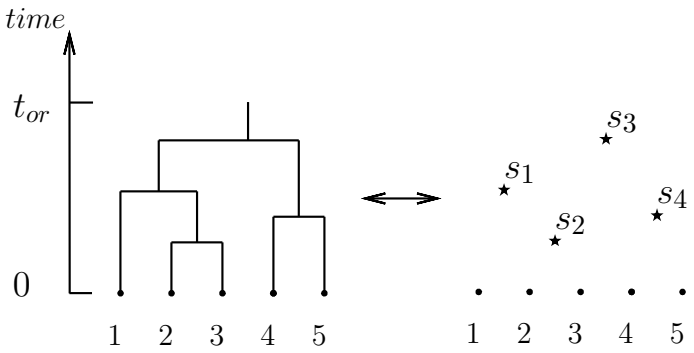


Figure 3.3: On the left, a tree with 5 species is displayed. On the right, we have the corresponding point process representation. The time t_{or} is the origin of the tree.

where t is the time of origin of the tree. The inverse of $F(s|t,\beta,\mu,n)$ is:

$$F^{-1}(s|t,\beta,\mu,n) = \frac{1}{\beta-\mu}\ln\left(\frac{\beta-\mu e^{-(\beta-\mu)t}-\mu(1-e^{-(\beta-\mu)t})s}{\beta-\mu e^{-(\beta-\mu)t}-\beta(1-e^{-(\beta-\mu)t})s}\right).$$

Recall that throughout this chapter, we assume a uniform prior for the time of origin of a tree. For this approach, we need the probability density of the time of origin of the tree, t, conditional on it having n species at the present. This distribution was derived in Gernhard (2008) for $\beta > \mu$:

$$Q(t|\beta,\mu,n) = \left(\frac{\beta(1-e^{-(\beta-\mu)t})}{\beta-\mu e^{-(\beta-\mu)t}}\right)^n.$$

The inverse of Q is

$$Q^{-1}(t|\beta,\mu,n) = \frac{1}{\beta-\mu}\ln\left(\frac{1-\frac{\mu}{\beta}t^{1/n}}{1-t^{1/n}}\right).$$

For $\beta = \mu$, the functions $F(s|t,\beta,\beta,n)$ and $Q(t|\beta,\beta,n)$ are established in

Aldous and Popovic (2005),

$$\begin{aligned} F(s|t,\beta,\beta,n) &= \frac{s}{1+\beta s}\frac{1+\beta t}{t} \\ F^{-1}(s|t,\beta,\beta,n) &= \frac{st}{1+\beta t(1-s)} \\ Q(t|\beta,\beta,n) &= \left(\frac{\beta t}{1+\beta t}\right)^n \\ Q^{-1}(t|\beta,\beta,n) &= \frac{1}{\beta(s^{-1/n}-1)} \end{aligned}$$

Combining these probability densities and the point process representation we obtain the following algorithm:

Constant rate birth-death approach

1. Sample $r_0,\ldots,r_{n-1}$ uniformly at random from $[0,1]$
2. Calculate the age of the tree, $t = Q^{-1}(r_0|\beta,\mu,n)$
3. Calculate the $n-1$ branching times, $s_i = F^{-1}(r_i|t,\beta,\mu,n), i = 1,\ldots,n-1$
4. Construct the tree from the point process representation
5. Repeat from step 1 until all samples have been obtained

The advantage of this method over *GSA* is that it is unnecessary to determine n^* and r. The disadvantage of this method is that it gives no information about extinct lineages (regardless of the value of μ). If this information is required, *GSA* must be used for sampling constant rate birth-death models. Finally note that a sample from the Yule model can be obtained by setting $\mu = 0$.

3.3 Comparison of the sampling approaches

We have shown that *SSA* is only appropriate for models without extinction where the time between speciation events is exponentially distributed with a rate parameter that depends only on the number of species that are extant. The two most popular models – the Yule and Coalescent – satisfy these

conditions and it is appropriate to sample from them using *SSA*. We speculate that the simplicity of *SSA* combined with its correctness for the two most popular models has resulted in its application to inappropriate models.

Existing approaches (such as *SSA*) are conceptually and computationally simpler than those introduced in this chapter, they have also been applied to many situations in existing studies for which they are inappropriate. It is therefore of great importance to consider how significantly the samples produced by the approaches differ. In situations where the difference is minimal it may be appropriate to use the simpler existing approaches to produce an approximate sample, if the difference is great it will be necessary to use more complicated approaches such as those presented here.

Lastly we note that throughout this section we have disregarded the root edge length. We therefore define the age of a sampled tree as the distance between the speciation event that created the second species and the leaves. This corresponds to realistic situations where it is often difficult to determine the length of the root edge.

3.3.1 Constant rate birth-death model

We begin by comparing *SSA* and *GSA* using a constant rate birth-death model – a simple extension of the Yule model that explicitly includes extinction – each species has the same probability per unit time of becoming extinct. The constant rate birth-death model includes two parameters – the speciation rate and the extinction rate – for our analysis it is sufficient to consider the ratio of these, hence we set the speciation rate to one. If the extinction rate is zero the model is equivalent to the Yule model. By increasing the extinction rate from zero to one the model becomes increasingly different from the Yule model and *SSA* will become increasingly inappropriate.

Figure 3.4 shows the expected age of the tree as a function of the extinction rate for samples of ten thousand trees produced by both sampling algorithms. When the extinction rate is zero the model is equivalent to the Yule model and the two approaches provide the same sample of speciation times. As the extinction rate is increased, the age of the trees sampled by *GSA* also increases as this effectively reduces the net speciation rate, resulting

in older trees.

We have shown that the absolute age of the tree differs for the two sampling approaches, however in some situations the relative timing of the speciation events may be all that matters. To investigate this feature we can consider lineage through time (LTT) plots which show the number of species present as a function of the age of the tree. When the number of species is log transformed the LTT plot should show a straight line with a deviation near the present (Nee et al., 1994; Harvey et al., 1994). Figure 3.5 shows the expectation of the LTT plot for an extinction rate of 0.95 from a sample of ten thousand trees produced using the two algorithms. There is a clear difference between *GSA* and *SSA*.

The slope near the origin of a log transformed LTT plot can be used to give an estimate of the net speciation rate. In Figure 3.5 we consider the difference between this slope for the two methods, as a function of the extinction rate. Interestingly around an extinction rate of 0.9 the bias switches from negative to positive.

Extinction rates have been estimated to be around 0.9 of the speciation rate (Magallon and Sanderson, 2001; Ricklefs, 2003). At this value the two sampling approaches differ significantly in the estimated age of the tree. For the relative timing of speciation events the result is not as clear, the severity (and direction) of the bias depends strongly on the extinction rate.

3.3.2 Tree shapes

The shape or topology of a tree is the structure obtained by disregarding the timing of speciation events (or equivalently the edge lengths). All memoryless models (including the constant rate birth-death model) produce trees with the same tree shape distribution. The reason for this is that there is nothing to differentiate between species, hence, regardless of the model, each species is always equally likely to be the one that undergoes the next speciation or extinction event. Furthermore since *SSA* does not distinguish between species it correctly samples the tree shape distribution for memoryless models.

SSA may incorrectly sample the tree shape distribution from models that feature a memory. For pure birth models, the mechanism behind this would

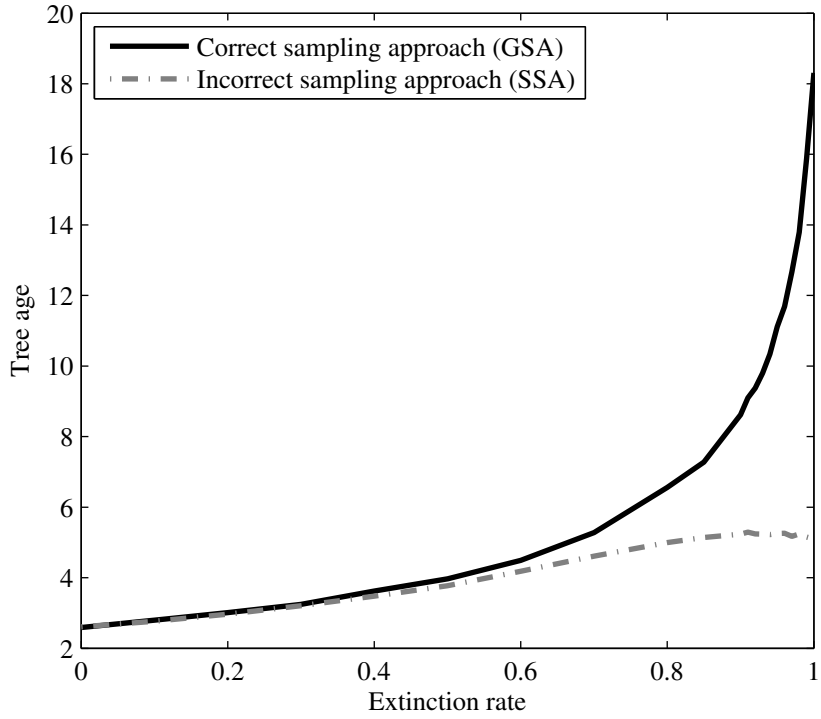


Figure 3.4: This figure shows the expected age for twenty-species trees sampled from the constant rate birth-death model as a function of the extinction risk. The speciation rate was set to one and 5000 trees were sampled for each extinction rate using *SSA* (dotted line) and *GSA* (solid line). The age of the trees sample by *GSA* increases as the extinction rate increases – this is because the net speciation rate is effectively reduced. *SSA* only considers the first time period during which n species existed, hence trees sampled using *SSA* do not exhibit the same age increase.

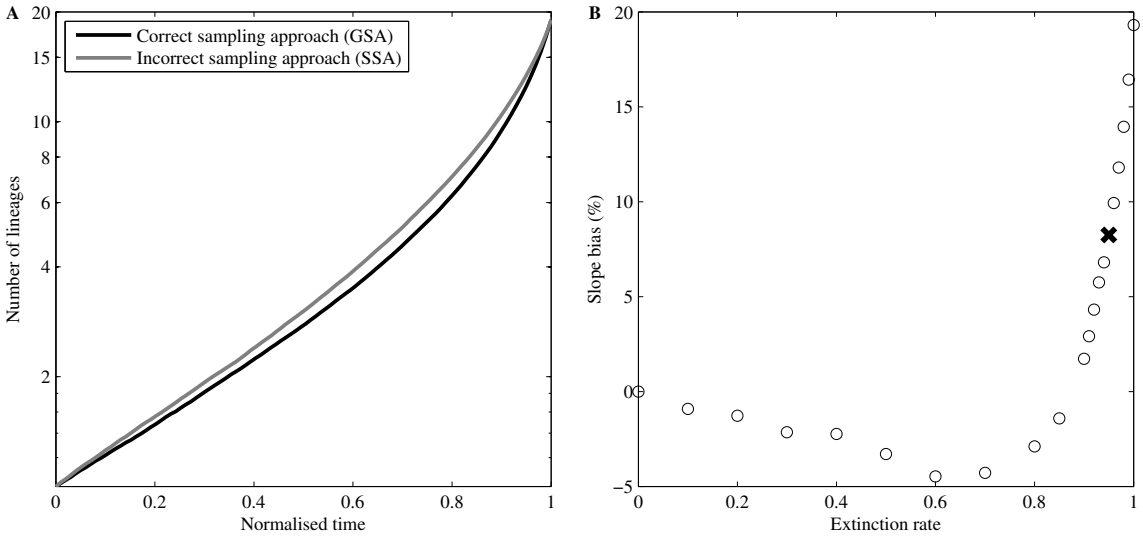


Figure 3.5: A: An expected lineage through time plot is shown here for five thousand, twenty-species trees sampled from a constant rate birth-death model using both *SSA* and *GSA*. The speciation rate was set to one and the extinction rate to 0.95. The trees have been rescaled to have age one – this removes the effect seen in Figure 3.4 and permits us to explore the relative speciation times of both samples. B: The initial slope in A gives an estimate of the net speciation rate. Here we depict the percentage deviation of the slope obtained using *SSA* to that obtained with *GSA*. The point corresponding to Panel A is marked.

require a correlation between the shape of a tree and the duration for which n species exist. This correlation is not explicit in any common models of which we are aware, but may exist implicitly; the strength of the correlation will determine the suitability of *SSA* to sample from a given model. We investigated two of the more common models with a memory (Heard, 1996; Blum and Francois, 2006) and found minimal bias in the tree shape distribution produced by *SSA*.

For other models *SSA* may introduce a more serious bias in the tree shape distribution. One of the most obvious cases is a model with extinction where the tree shape distribution changes over time – as we have seen *SSA* produces trees that are too young, hence the tree shape distribution would be sampled too early.

3.3.3 Incomplete taxon sampling

The most common approach for incomplete taxon samples a tree containing the expected true number of species, $\hat{m}$, and then randomly deletes $n - \hat{m}$ of these species. Here we have provided an extension to *GSA* that considers a range of possible true tree sizes and samples these accordingly. We applied this method to the constant rate birth-death model and found that the sampled trees differed negligibly from those obtained using the conventional approach. There are two main issues with the conventional approach, in this section we illustrate why each issue results in only a negligible bias:

Issue 1: Consider the constant rate birth-death model, Figure 3.6 shows how the expected age of a ten-species tree suffering from incomplete taxon sampling increases as a function of the true tree size. It is important to note that this is near-linear; in unpublished results Tanja Gernhard has shown that for the constant rate birth-death model the relationship is linear when the extinction rate is one, and becomes slightly non-linear as the extinction rate is decreased. If this relationship were perfectly linear and the true number of species were known, sampling a tree with m species and deleting $n-m$ species would give a correct sample. For this model the deviation from linearity seems sufficiently small to be irrelevant for most purposes.

Issue 2: Given a probability of sampling each species (s), a naive method

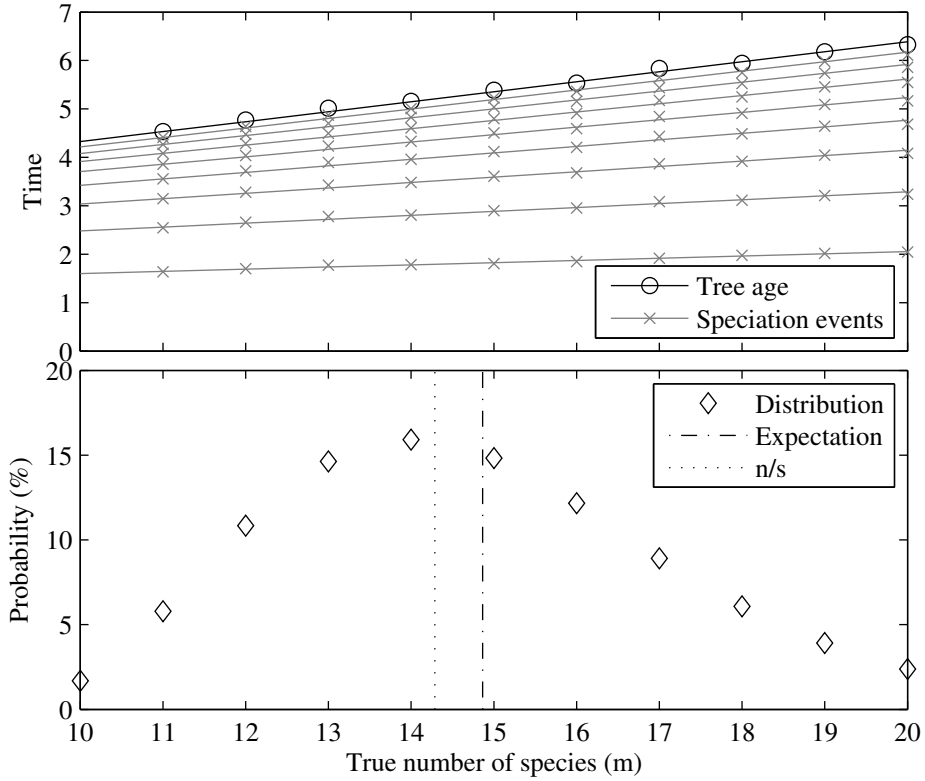


Figure 3.6: Top panel: The black circles show the expected age of a ten-species tree that has been sampled by constructing a m species tree and deleting $m - 10$ species. Five thousand samples were taken for each value of m using the constant rate birth-death model with speciation rate 1 and extinction rate 0.9. The gray crosses show the expected time of the speciation events in the same situation. The lines are linear least-squares fits to these points, demonstrating that the relationships are near-linear. Bottom panel: The diamonds show the probability distribution of the true tree size, m, as calculated from Equation 3.5 for a sampling probability of $s = 0.7$. Also depicted are the expectation of this distribution (about 14.8) and a simple estimate of this – n/s (about 14.3).

for calculating the expected number of species would be $\hat{m} = n/s$. In Figure 3.6 we show the distribution of the true tree size as calculated using Equation 3.5 for $s = 0.7$, due to the asymmetry of this distribution, its expecation exceeds n/s. In this example the difference between these expectations is about 0.5, this will result in a small bias towards younger trees.

For the constant rate birth-death model, the bias introduced by using a simplistic incomplete sampling method is insignificant in contrast with uncertainty regarding the true number of species. For other models it may be necessary to use the approach the full sampling approach outlined here. This will particularly be the case for models that exhibit a strong non-linearity in the expected age curve shown in Figure 3.6.

3.4 Concluding comments

When exploring evolutionary models, analytic results are preferable to simulation studies because of the smaller computational burden and greater insight they provide. However analytic results may be difficult to obtain and simulation studies may answer questions more quickly – once a result has been confirmed by simulation studies an analytic approach can be pursued with extra confidence.

Simulation studies have an inherent danger – it is extremely easy to simulate trees using a given model, however understanding what distribution these trees come from can be difficult. This makes it easy to proceed with a (possibly incorrect) method and therefore sample of trees. This is particularly problematic with more complicated evolutionary models where seemingly intuitive methods of simulating trees (such as *SSA*) often sample from undesirable and unrealistic probability distributions.

We have shown that a commonly used sampling approach is appropriate for two of the most common evolutionary models – the Yule and Coalescent. However this approach is inappropriate for many other models to which it has been applied. For the constant rate birth-death model, *SSA* produces a strong bias in the age of the tree and the relative timing of speciation events. It does not produce a bias in the tree shape distribution. Further, for the birth-death model, the common approach for incorporating incomplete taxon

sampling seems adequate for most applications. More complex models with certain characteristics as discussed in this chapter may result in stronger biases of any of these attributes of a sampled tree.

We suggest that for some of the studies that have sampled trees using *PhyloGen* and *SSA*, it would have been more appropriate to use our presented methods. It should be noted that in the cited studies the sampled trees were only one part of a complicated process (eg. to generate a data set for testing a tree construction method) and it is unlikely that the results would have been significantly effected by the chosen sampling method. For studies explicitly comparing speciation times in trees or sampling from more complicated models the distinction between these distributions will become crucial.

The methods presented here have been implemented in the Perl module Bio::Phylo. It is easy to apply this implementation of our methods to other appropriate evolutionary models. A simple GUI for these algorithms, TreeSample, is also available for users unfamiliar with Perl. TreeSample has built in support for the Yule model and constant rate birth-death models and is extendable to permit sampling from additional models.

We hope that this chapter helps clarify some of the issues about sampling trees from evolutionary models and that the software we have created will be of use for future simulation studies.

Chapter IV

Artificial trees are too balanced!

In this chapter we consider a simple property of phylogenetic trees – tree imbalance. This is of particular interest as trees derived from 'real' data and trees produced by simple memoryless models such as the Yule model have different distributions of tree imbalance. More complicated models that attempt to address this issue have been proposed, however these models have unsatisfying biological interpretations. Here we present an alternative model that has a simple biological interpretation and matches 'real' trees. This model is fitted to trees from several data sources producing a surprising fit.

4.1 Tree imbalance

Phylogenetic tree imbalance is a measure of the asymmetry in a tree, disregarding leaf labels and edge lengths. From a cursory examination of the trees in Figure 4.1 it is clear that D is most imbalanced (a caterpillar) and A is as balanced as possible – but how do B and C compare?

To formalise the notion of imbalance many indices have been introduced. The most widely used index is I_c (Colless, 1982) which considers the difference in the size of the two daughter trees descendant from each internal node and adds this difference across the tree:

$$I_c = \frac{\sum_{v \in \mathring{V}} |\tau_{v,a}| - |\tau_{v,b}|}{\frac{(n-1)(n-2)}{2}},$$

where $\mathring{V}$ is the set of interior nodes and $\tau_{v,a}$ and $\tau_{v,b}$ are the two trees descendant from a node v. The denominator normalises I_c such that $I_c = 1$ for a completely unbalanced tree (a caterpillar) and $I_c = 0$ for a perfectly balanced tree (note that this is only possible for trees where the number of leaves is a

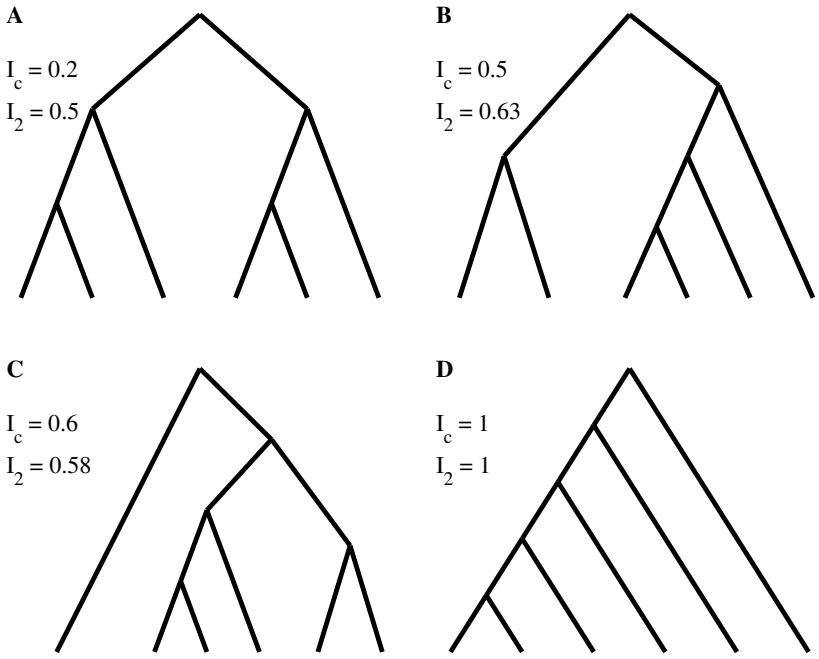


Figure 4.1: Four tree shapes are shown along with their values for two common tree imbalance statistics. Note that whether B is more imbalanced than C depends on the imbalance statistic that is used.

power of two). It should be noted that I_c places greater emphasis on nodes near the root as these have larger descendant trees and can therefore make a greater contribution to I_c. This was observed by Kirkpatrick and Slatkin (1993) who introduced an alternative index I_2 that weights all internal nodes equally:

$$I_2 = \frac{\sum_{v \in \mathring{V}} \frac{|\tau_{v,a}| - |\tau_{v,b}|}{C_v - 2}}{n-2},$$

where C_v is the number of species descendant from node v. Many further indices are possible (see Kirkpatrick and Slatkin (1993) and Matsen (2006) for a discussion), however the results in this section are independent of the index used.

Tree imbalance (regardless of how it is measured) is a property of tree shape and can be considered a projection from the multi-dimensional tree space to a one dimensional index (Matsen, 2006). As discussed in chapter 3 all memoryless models have the same tree shape distribution and consequently the same distribution of tree imbalance. We will adopt the convention found in the literature and refer to this distribution as the Yule tree shape or imbalance distribution.

It has been widely observed that trees derived from 'real' data are more imbalanced than those from a Yule models (Mooers and Heard, 1997; Aldous, 2001). Realistic models should therefore feature a memory, and several models have been introduced that can reproduce the imbalance in 'real' trees. However the parameter values required by these models have extreme biological implications.

4.1.1 Proportional to distinguishable arrangements model

We now describe a model that is often considered at the 'opposite' end of the tree balance spectrum to the Yule model – the proportional to distinguishable arrangements model (PDA model; Aldous (1996, 2001); Semple and Steel (2003)). The PDA model states that each distinguishable labelled tree should be equally probable. Consider the two possible tree shapes with four species. The unbalanced tree shape has $4 \times 3 = 12$ distinguishable leaf labellings as the two species not contained in the cherry can be uniquely identified. The

balanced tree shape only has 3 possible labellings as the two species contained in each of the cherries are not distinguishable. Under the PDA model the balanced tree shape therefore has a probability of 0.2 whereas under the Yule model it has a probability of 1/3.

The PDA model predicts a distribution of tree shapes that is more imbalanced than those predicted by the Yule model. The Yule and the PDA can be considered at opposite ends of the tree balance spectrum with 'real trees' somewhere in between (Blum and Francois, 2006; Pinelis, 2003; Aldous, 1996). The PDA model does not have a simple biological explanation, however some models that include the PDA distribution as a special case have been developed.

4.2 Existing models

Several models have been proposed that can give a tree balance distribution resembling that of real trees, however these models lack biological plausibility. Pinelis (2003) introduced a model which spans a range of possible tree balance distributions, but the corresponding evolutionary interpretation is unclear, and the model has seen little subsequent use.

Several models that feature randomly 'evolving' speciation rates have been proposed and it has been shown that they can produce trees that are sufficiently imbalanced (Blum and Francois, 2006; Heard, 1996; Heard and Mooers, 2002). However these models require speciation rates to 'evolve' at an unrealistic pace, this was shown in (Heard, 1996) for their model and here we show that the beta-binomial model from Blum and Francois (2006) also suffers from this problem.

4.2.1 The beta-binomial model

The Beta Binomial (BB) model (Blum and Francois, 2006) provides a good fit to the 'real' trees they investigated. The BB model is based on a speciation rate, λ, that randomly evolves along lineages. More specifically when a species with speciation rate λ undergoes a speciation event, a number p is drawn from a beta distribution and the new species are assigned speciation rates of $p\lambda$ and $(1-p)\lambda$. Here we show that the parameter values Blum and

Francois obtained imply extremely rapid changes in λ.

In the BB model the parameter p is distributed according to a beta distribution:

$$p \sim \text{Beta}(\alpha + 1, \alpha + 1), \qquad \alpha > -1,$$

where α is the parameter that can be varied to fit the model to the data. Blum and Francois suggested a value for α of -0.58. Simple analyses (not presented here) of the datasets described at the end of this chapter also suggest values for α between -0.3 and -0.65 (depending on the data set and the tree size). Panel A in Figure 4.2 shows the probability distribution of p for a range of α values. Consider those α values that correspond to real data sets, their probability densities have maxima at $p = 0$ and $p = 1$; these high and low values of p correspond to the situation where the new species have very different speciation rates.

Panel B in Figure 4.2 shows the cumulative probability distribution for the lower value of p obtained after a speciation event. For example if $\alpha = -0.58$ this shows that in over 40% of events the speciation rates differ by a factor of ten, quite extreme from a biological perspective. When α is greater than zero the BB model has a more realistic biological interpretation. Consider Figure 4.2, if $\alpha = 2$ for example, only about 1% of speciation events are expected to result in species with a speciation rate that differs by a factor of ten.

The BB model suggests that frequent speciation bursts involving a single species will occur ie. after a speciation event one of the new taxa is likely to rapidly undergo another speciation event, then one of those new taxa is again likely to undergo rapid speciation and so on whilst all other taxa (including the other taxon 'born' in these speciation events) are unlikely to undergo speciation until much later. In contrast the model we will introduce implies that after speciation all new species have a heightened probability of speciating followed by a near constant speciation rate similar to the Yule model. This has the additional advantage of simplicity as it considers all taxa to be equal and does not require speciation rates to be inherited along lineages.

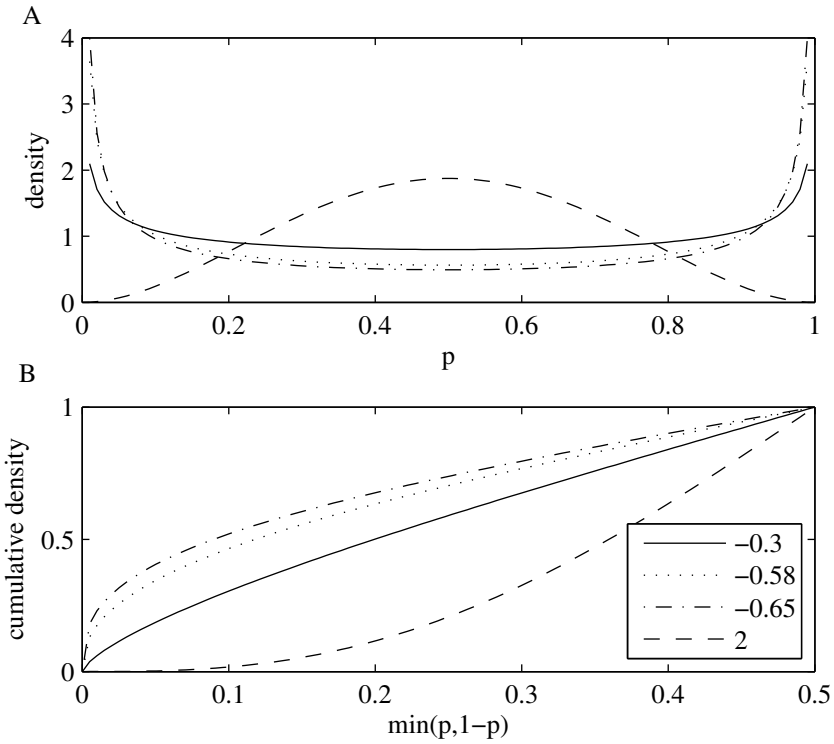


Figure 4.2: A: The beta distribution for values of α that are of interest, note the peaks in the distribution at zero and one for $\alpha < 0$. The value $\alpha = 2$ is unrealistic and included only for comparison. B: The cumulative distribution for the lower speciation rate, this highlights the rapid changes in the speciation rate that will occur in the BB model with realistic α values.

4.3 The Weibull Bellman Harris model

The model we propose here is a Bellman-Harris (BH) model (as introduced in chapter 2), where the time to the speciation event on a given lineage is sampled from a Weibull distribution; we refer to this model as the Weibull Bellman Harris (WBH) model. We chose to investigate the Weibull distribution for several reasons. Firstly it is a life time distribution, these are typically used to model the time to failure or death of machinery or organisms and are of particular interest as they can be derived from principles that may also apply to an evolutionary process. Secondly we will show that a BH model with the Weibull distribution includes both the Yule and PDA (Aldous, 1996) models as special cases. Inclusion of these models is useful as they are frequently used and often considered at opposite ends of the tree balance spectrum. BH models with different distributions may also be of interest and we encourage the investigation of such models.

The Weibull distribution is suitable for modelling situations where there are many parts with the same failure distribution and we want to know the time to the first failure (Nelson, 1982). This is a 'chain model' where the strength of the chain is determined by the strength of the weakest link. In the context of an evolutionary model the individual parts, for example, could be considered different populations of the same taxon each of which has the same probability distribution of undergoing a speciation event.

By choosing an appropriate time scale (non-dimensionalising) and requiring that all branch lengths are possible we can restrict ourselves to considering the one parameter Weibull distribution, which is given by $g(t) = \beta t^{\beta-1} e^{-t^\beta}$. An important characteristic of this distribution is the rate at which those taxa that have not yet speciated undergo speciation events. This is commonly known as the failure rate although in our context we refer to it as the speciation rate; mathematically it is given by $\lambda(t) = g(t)/\int_t^\infty g(t)dt = \beta t^{\beta-1}$. Figure 4.3 shows the Weibull distribution ($g(t)$) and the associated speciation rate ($\lambda(t)$) for various β values. The WBH process corresponds to the Yule and PDA models at $\beta = 1$ and $\beta = 0$ respectively. In terms of tree balance real trees lie somewhere between these two (Blum and Francois, 2006; Pinelis, 2003; Aldous, 1996) consequently we focus on the range $0 < \beta \leq 1$.

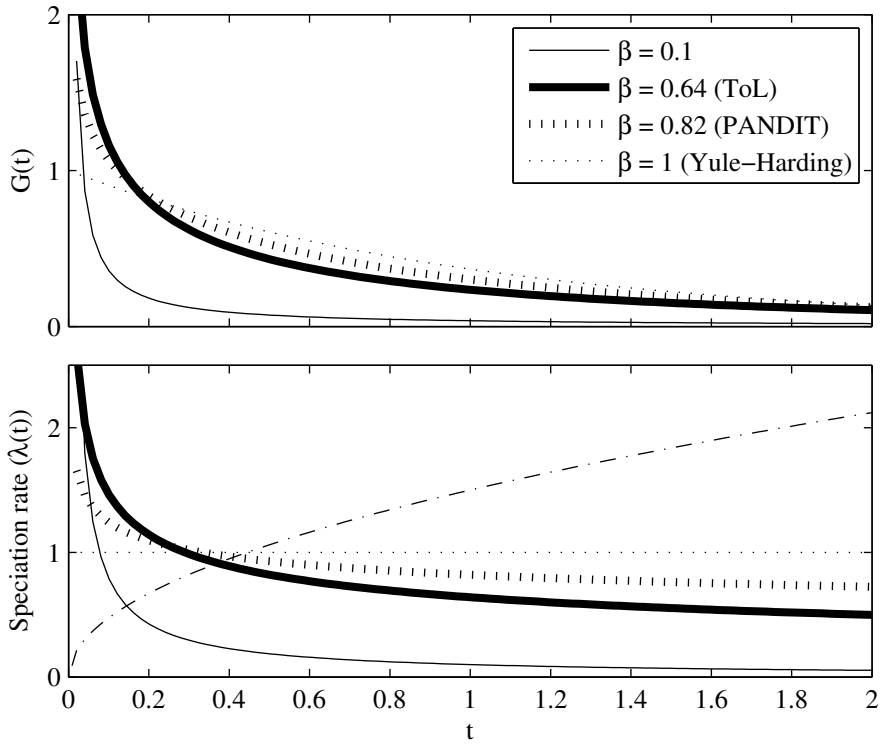


Figure 4.3: The top panel shows the Weibull probability density for a range of β values. The bottom panel shows the speciation rate – the rate at which species that have not yet speciated undergo speciation.

From Figure 4.3 and the speciation rate equation it is evident that the speciation rate decreases with time when $\beta < 1$. This is consistent with a range of biological scenarios which result in bursts of speciation (Steel and McKenzie, 2001). For the parameter values we later obtain (β between 0.6 and 0.8) the speciation rate is initially high, corresponding to speciation bursts, followed by a near constant speciation rate corresponding to the situation described by the Yule model.

4.3.1 Limit results

Here we prove that the WBH model is equivalent to the Yule and PDA models when $\beta = 1$ and in the limit as $\beta \to 0$ respectively. Recall from chapter 2 that $p(\tau|t)$ is the probability that a tree of age t will have a given shape (τ) and that $g(t)$ is the probability density for the time between speciation events on a lineage. Here we make the dependence of these quantities on β explicit by referring to them as $p(\tau|\beta, t)$ and $g(t|\beta)$.

Theorem 5. *In the limit as $\beta \to 0$ the WBH model induces the PDA probability distribution on tree shapes for any $t > 0$. Furthermore for a tree shape, τ , with n leaves we have:*

$$p(\tau) := \lim_{\beta \to 0} p(\tau|\beta, t) = 2^{u(\tau)} e^{-n} \left(1 - e^{-1}\right)^{n-1}, \qquad \infty > t > 0, \tag{4.1}$$

where $u(\tau)$ is the number of unbalanced internal vertices of of τ.

Proof. Note that the Weibull probability distribution has the following properties (for $t > 0$):

$$\lim_{\beta \to 0} g(t|\beta) = \lim_{\beta \to 0} \beta t^{\beta - 1} e^{-t^\beta} = 0 \tag{4.2}$$

$$\lim_{\beta \to 0} \int_t^\infty g(u|\beta) du = \lim_{\beta \to 0} e^{-t^\beta} = e^{-1} \tag{4.3}$$

From chapter 2 we have the following (Equation 2.1, where the dependence on β has been made explicit):

$$p(\tau|\beta,t) = \begin{cases} \nu_\tau \int_0^t g(\lambda_r|\beta)p(\tau_a|\beta,t-\lambda_r)p(\tau_b|\beta,t-\lambda_r)d\lambda_r & |\tau|>1 \\ 1-\int_0^t g(u|\beta)du & |\tau|=1, \end{cases} \tag{4.4}$$

where ν_τ equals two if τ_a and τ_b are different and one if they are equal. We now consider the limit of $p(\tau|\beta,t)$ as $\beta \to 0$.

Firstly, if $|\tau|=1$ we have:

$$\begin{aligned} \lim_{\beta\to 0} p(\tau|\beta,t) &= \lim_{\beta\to 0}\left(1-\int_0^t g(u|\beta)du\right) \\ &= e^{-1}. \end{aligned}$$

If $|\tau|>1$ we have:

$$\begin{aligned} \lim_{\beta\to 0} p(\tau|\beta,t) &= \nu_\tau \lim_{\beta\to 0}\int_0^\epsilon g(\lambda_r|\beta)p(\tau_a|\beta,t-\lambda_r)p(\tau_b|\beta,t-\lambda_r)d\lambda_r + \\ &\quad \nu_\tau \lim_{\beta\to 0}\int_\epsilon^t g(\lambda_r|\beta)p(\tau_a|\beta,t-\lambda_r)p(\tau_b|\beta,t-\lambda_r)d\lambda_r, \end{aligned} \tag{4.5}$$

where ϵ is arbitrarily small and lies between 0 and t. Since $p(\tau|\beta,t-u)$ is a probability, it is strictly less than unity and using Equation 4.2 we have the following bound on the last integral in Equation 4.5:

$$\nu_\tau \lim_{\beta\to 0}\int_\epsilon^t g(\lambda_r|\beta)p(\tau_a|\beta,t-\lambda_r)p(\tau_b|\beta,t-\lambda_r)d\lambda_r \le \lim_{\beta\to 0}\int_\epsilon^t g(u|\beta)du = 0, \tag{4.6}$$

as the integrand is strictly positive this is a strict equality. Using Equations 4.3 and 4.6 and the fact that ϵ can be arbitrarily small, Equation 4.5 becomes:

$$\begin{aligned} \lim_{\beta\to 0} p(\tau|\beta,t) &= \nu_\tau \lim_{\beta\to 0}\int_0^\epsilon g(\lambda_r|\beta)p(\tau_a|\beta,t-\lambda_r)p(\tau_b|\beta,t-\lambda_r)d\lambda_r \\ &= \nu_\tau \lim_{\beta\to 0} p(\tau_a|\beta,t)p(\tau_b|\beta,t)\lim_{\beta\to 0}\int_0^\epsilon g(\lambda_r|\beta d\lambda_r \\ &= \nu_\tau p(\tau_a)p(\tau_b)(1-e^{-1}). \end{aligned}$$

In summary the limit of Equation 4.4 is:

$$\lim_{\beta \to 0} p(\tau|\beta, t) = \begin{cases} \nu_\tau p(\tau_a) p(\tau_b)(1 - e^{-1}) & |\tau| > 1 \\ e^{-1} & |\tau| = 1. \end{cases} \tag{4.7}$$

We now prove Equation 4.1 using induction. For a tree with one leaf ($n = 1$) we have already shown that Equation 4.1 holds. Now assume Equation 4.1 holds for all trees with fewer than n leaves and let τ be a tree shape with n leaves and subtrees τ_a and τ_b attached to the root. The two subtrees have n_a and n_b leaves with $n_a \leq n_b$ and $n_a + n_b = n$. Using Equations 4.7 and 4.1 we have:

$$\begin{aligned} \lim_{\beta \to 0} p(\tau|\beta, t) &= \nu_\tau p(\tau_a) p(\tau_b)(1 - e^{-1}) && (4.8) \\ &= \nu_\tau 2^{u(\tau_a)+u(\tau_b)} e^{-n_a - n_b} \left(1 - e^{-1}\right)^{(n_a-1)+(n_b-1)+1} && (4.9) \\ &= 2^{u(\tau)} e^{-n} \left(1 - e^{-1}\right)^{n-1}, && (4.10) \end{aligned}$$

as required. The PDA probability distribution for tree shapes of a given size is known to be proportional to $2^{u(\tau)}$ (Steel and McKenzie, 2001), hence conditioning $p(\tau)$ on n will result in the PDA probability distribution.

Theorem 6. *For $\beta = 1$ the Weibull model induces the same distribution on tree shapes as the Yule model.*

Proof. For $\beta = 1$ the Weibull probability distribution becomes $g(x) = e^{-x}$, which is simply the exponential distribution with the scaling parameter removed (through non-dimensionalisation). By definition this is the Yule model.

4.3.2 Extinction

Like many other evolutionary models the WBH model does not explicitly include extinction, instead the probability distribution should be interpreted as the probability distribution of a speciation event occurring where both new species survive to the present, thus implicitly including extinction. This avoids the problem of including extinction and is justifiable as we have made no attempt to derive the Weibull distribution from biological processes (we

have merely considered the biological implications it suggests). The problem with this interpretation is that the probability of a species surviving from speciation to the present should depend on its age. This suggests that the probability distribution for a successful speciation event should change as a function of absolute time, which it does not in the WBH model.

Here we consider treating the Weibull distribution, $g(t|\beta)$, as giving the probability distribution of any speciation event occurring and consider extinction a random process such that a species has probability, p, of surviving from 'birth' to the present. The probability that the first successful speciation event occurs at t is therefore found by considering all the possible number of unsuccessful events that have taken place by that time. Given that i 'unsuccessful' speciation events have occurred previously, the probability of a speciation event occurring at time t is:

$$\hat{g}_i(t|\beta) = \begin{cases} g(t|\beta) & i = 0 \\ \int_0^t g(u|\beta)\hat{g}_{i-1}(t-u)du & i > 0. \end{cases}$$

The probability distribution of a successful speciation event at time t is therefore:

$$\acute{g}(t|\beta) = p\sum_{i=0}^{\infty}(1-p)^i\hat{g}_i(t|\beta). \tag{4.11}$$

A Weibull distribution was fitted to $\acute{g}$, achieving a seemingly exact fit. To achieve this fit the scaling parameter in the Weibull distribution had to be allowed to vary from unity, this is to be expected as introducing extinctions changes the characteristic time scale. The exact fit seems universal over all β and p values investigated, however a proof remains open.

Figure 4.4 shows the relationship between the survival probability, p, and the new β parameter from the resultant distribution $\acute{g}(t)$. Note that β increases with a decreasing survival probability. The survival probability, p, should be higher for more recent times, hence β should decrease in a tree from the root to the leaves. It would therefore be expected that smaller trees should correspond to lower β values than larger trees.

Equation 4.11 also provides a compact alternative proof that extinction

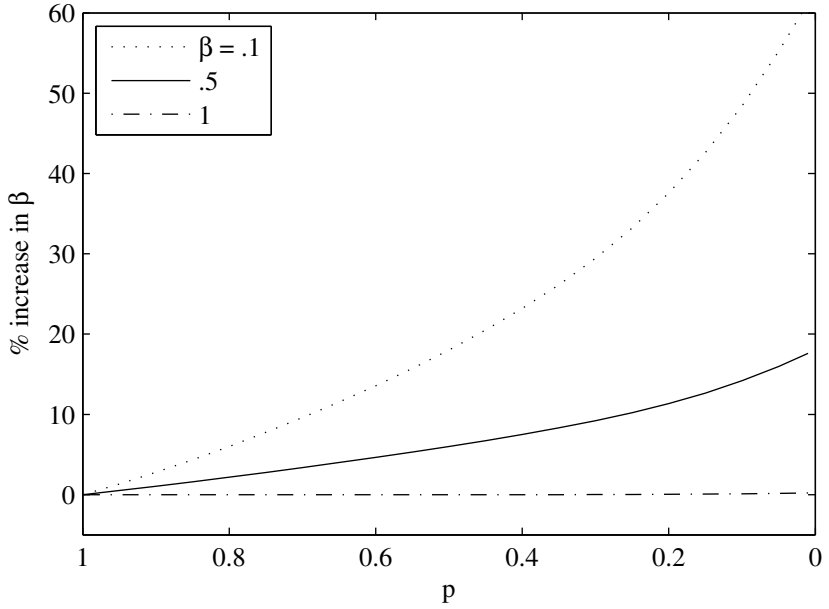


Figure 4.4: The increase in β as a function of the survival probability of a species, p, for a range of β values.

can be included in the Yule model by using a net speciation rate, using straightforward induction for the Yule model it can be established that:

$$\hat{g}_i(t) = \frac{1}{i!}t^i e^{-t}$$

Substitution in Equation 4.11 yields:

$$\begin{aligned} \acute{g}(t) &= p\sum_{i=0}^{\infty}\frac{(1-p)^i}{i!}t^i e^{-t} \\ &= pe^{-t}\sum_{i=0}^{\infty}\frac{(1-p)^i}{i!}t^i e^{-t} \\ &= pe^{-pt}, \end{aligned}$$

an exponential distribution with scaling parameter $1/p$, as required. It should be noted that p varies with time, hence the net speciation rate used in the Yule model should also vary with time.

4.4 Fit to data

To test the WBH model four online databases containing phylogenetic trees were considered TreeBASE (Sanderson et al., 1994), Tree of Life (ToL) (Maddison and Schulz, 2006), PANDIT (Whelan et al., 2006) and HOVERGEN (Duret et al., 1994). The last three contain trees respectively corresponding to species, protein domains and vertebrate genes, whereas TreeBASE contains any trees authors wish to submit. For simplicity we continue to refer to the individuals in these trees as species.

4.4.1 Real tree imbalance distribution

For each tree size, n, all interior edges in all trees in the database were examined and all trees corresponding to interior edges with n leaves were extracted. This approach was taken (as opposed to simply selecting trees in the database with n leaves) to maximise the data available for the analysis, the necessity of this is particularly evident with the ToL database which contains only one tree.

The databases contain multifurcations (where one ancestral taxon splits into more than two new taxa), each extracted tree containing multifurcations was equally attributed to each tree shape that refines it. For each family in the PANDIT database three trees are provided based on different alignments, the trees corresponding to each type of alignment were treated as different datasets. The results obtained from two of these datasets (aa and aa-restricted) were nearly identical hence we only show results for the aa and dna alignments.

The protein trees in PANDIT and HOVERGEN are highly correlated. When proteins from several species are in a single tree then the that tree will to a certain extent be constrained by the underlying species tree. Since these species may be contained in several of the protein trees, the protein trees will not be independent. To counter this only those extracted trees containing proteins from a single species were considered – these contain modern protein evolution.

4.4.2 Comparing the WBH model and real trees

Approach 1 in chapter 2 was used to calculate the probabilities of tree shapes occurring under the WBH model. The resulting nested integrals were solved numerically. Figure 4.5 depicts the probabilities that the WBH model places on obtaining various tree shapes for small trees. These probabilities are shown for a range of the WBH model's β parameter from the PDA to the Yule model. Corresponding tree shape distributions were extracted from the databases as outlined in the methods and are shown in this figure at the β value where the WBH model provides an optimal fit. The perfect fit for trees of size four is expected as there are zero degrees of freedom, however the fit for the larger tree sizes is remarkable especially for the PANDIT and HOVERGEN data sets.

4.5 Concluding comments

In this chapter we have introduced a new biologically motivated model – the WBH model. This model spans a range of tree imbalance distributions from the Yule to the PDA model (and beyond), within this range it provides an excellent fit to the trees obtained from several large online databases. The WBH model provides a simpler and more realistic biological interpretation than previous models that have achieved this. It simply stipulates that the speciation rate decreases with the age of a species, this is consistent with bursts of speciation.

A good fit of the WBH model to the trees from four databases was obtained. The tree databases used in this study are massive, freely available and easy to obtain. Unfortunately these tree database also have some shortcomings. TreeBASE includes trees from a broad variety of sources and there is no quality control. The ToL is highly incomplete, although we have tried to correct for this. Lastly PANDIT and HOVERGEN contain trees that are highly correlated. Compilation of datasets that have undergone a higher level of quality control will provide more accurate tests of the WBH (and other) models.

In this study we examined the fit of the WBH model to tree shape distributions. The timing of speciation events was not considered, this would

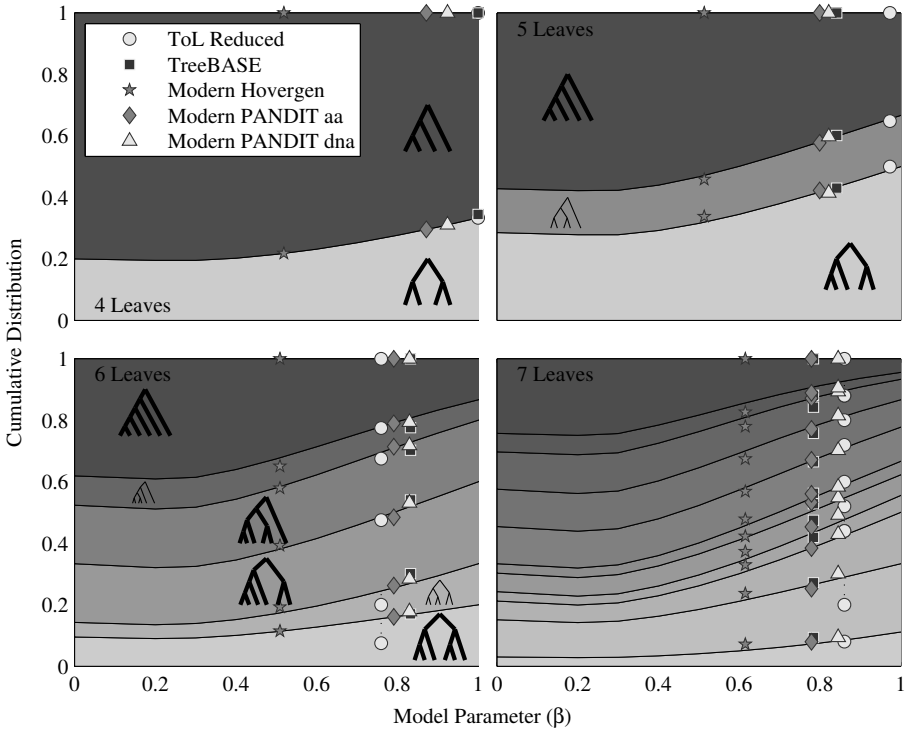


Figure 4.5: Cumulative probability distributions as obtained from the WBH model for small trees over a range of β values from 0 to 1, which respectively correspond to the PDA and Yule model. Past studies suggest that real trees should lie somewhere in this range. Trees are arranged from top to bottom from theleast balanced tree shape to the most balanced tree shape. To clarify the interpretation consider trees with four leaves (taxa), if the model parameter (β) is 0.8 the cumulative probabilities of the tree shapes are roughly 0.3 and 1; this indicates that out of the trees with four species generated by the WBH model 30% are as balanced as possible and 70% have the remaining tree shape. The distributions observed in actual datasets are plotted at the β value for which the best fit was obtained.

complicate the analysis and create further problems with the sets of trees that were used – TreeBASE and ToL do not contain edge lengths. This would be a worthwhile topic for future analysis.

Part II

Biodiversity Conservation

Chapter V

Introduction

It is commonly believed that we are presently undergoing the first mass extinction event since the Cretaceous-Tertiary event (which was responsible for the demise of the dinosaurs). Our current extinction event may be the most rapid ever (Louis Harris & Associaties, 1998), with the present extinction rate estimated to be between 100 and 1000 times the background extinction rate (Lawton and May, 1995). IUCN (2007) has documented the extinction of 785 species and 39% of the 1.6 million species listed on the IUCN Red List are threatened. The number of documented extinctions is a vast underestimate as most cases go unnoticed, the real rate may be as high as 140,000 species per year (Pimm et al., 1995).

This incredible loss of biodiversity undermines ecosystem stability and threatens their ability to supply goods and services vital to humans (Secretariat of the Convention on Biological Diversity, 2006). It is of crucial importance to reduce the loss of biodiversity through conservation. Given the limited resources available for conservation we should ensure that these resources are spent in a way that best achieves our goals. The question is: what is the ultimate goal of conservation?

Ross Crozier summarizes the rationales for conserving biodiversity into three categories: "moral (other species have a right to exist), esthetic (species are like works of art, and it would be foolish to destroy them), and utilitarian (humans derive material benefit from the existence of other species)" (Crozier, 1997); these motivations are further explored in Norton (1987). Given unlimited resources for conservation all three motivations dictate the same action - conserving all taxa. In a realistic setting where there are limited resources for conservation the taxa must be prioritized in some manner. In this case the three categories of motivation may dictate different prioriti-

sations.

If conservation is motivated by moral considerations, as many taxa as possible should be conserved. A conservation scheme should therefore allocate its resources so that the net survival increase of all taxa is as high as possible.

If the motivation for conservation is utilitarian, the distinctiveness of the remaining taxa is of great importance. For example, protecting the sole remaining taxon from a clade has greater utilitarian benefits than protecting a taxon from a well represented clade as the former has greater unique genetic potential for further evolution and bio-prospecting (Crozier, 1997).

Lastly if conservation is motivated by esthetic reasons the role that distinctiveness should play is dependent on the uncertain definition of esthetic value. However given the choice of saving either a taxon from a well represented clade or a taxon that is the 'last of its kind' it seems difficult to find a general justification for not choosing the latter.

Most biodiversity conservation approaches aim to conserve as many taxa as possible (Gaston, 1996), but the reasons used to motivate conservation are often utilitarian in nature (eg. chapter 1 of Pullin (2002)) and should therefore take taxon distinctiveness into account. In Figure 5.1 it is clear that Daubentonia is the most distinctive species. But for example, identifying the four most important species is not as straightforward and necessitates a formal measure of distinctiveness / biodiversity.

In this part of the thesis we will consider two types of biodiversity measures that utilise a phylogenetic tree. The first type is Phylogenetic Diversity (PD; Faith (1992), chapter 6), a measure of the biodiversity of a set of species. PD has been widely used as a biodiversity measurement but is unsuitable for many current management methodologies which require species to be ranked in order of importance. Due to this problem we also consider a second type of biodiversity measure – species specific indices (chapter 10). Using the phylogeny species specific indices allocate each species a unique index value that gives some indication of that species' contribution to biodiversity. Prioritising species using this method will result in a suboptimal set of species being selected, however we show that some indices are still capable of producing good solutions.

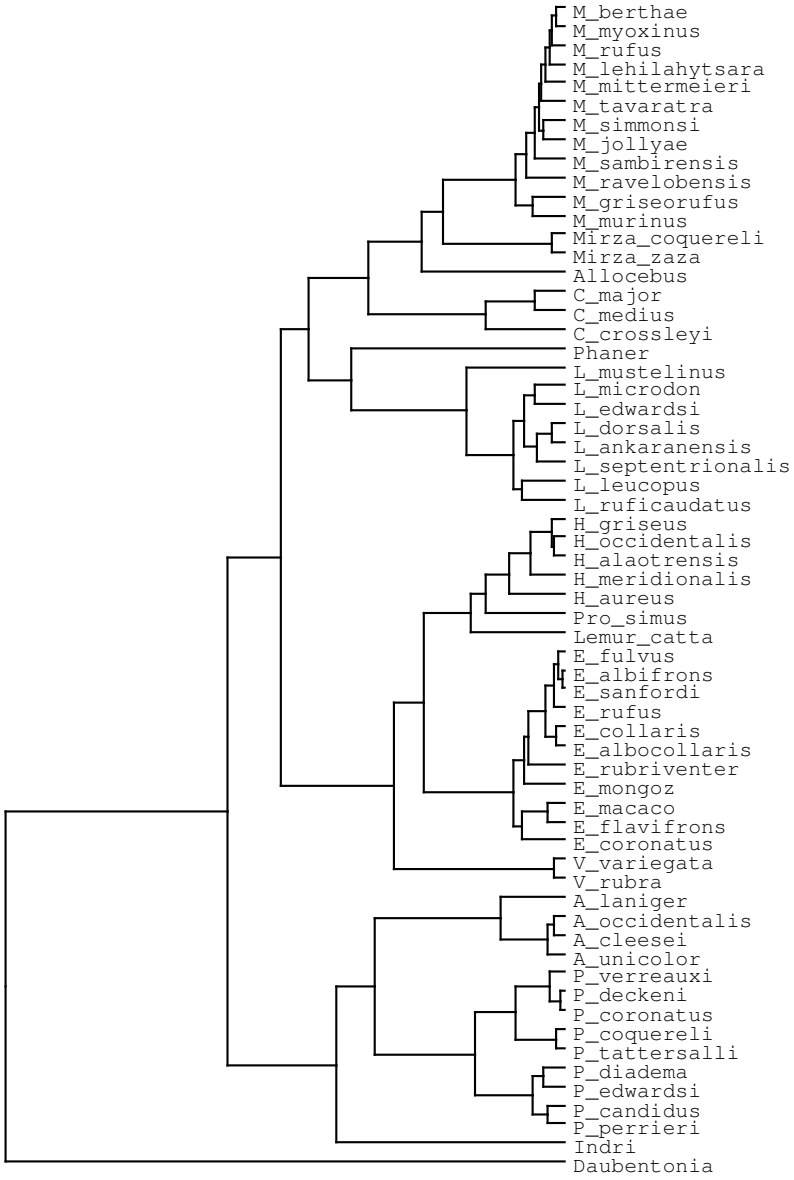


Figure 5.1: The phylogenetic tree for the Madagascan lemurs from Hartmann et al. (b).

Generally species will have different survival probabilities and different conservation costs. A framework that takes this into account and aims to maximise the expected PD of the surviving species is referred to as the Noah's Ark Problem (NAP; Weitzman (1998), chapter 7). Unfortunately finding optimal solutions to the NAP is a computationally intensive problem. In chapter 7 algorithms for several restricted instances of the NAP are provided.

A further problem with the NAP is that the input parameters (survival probabilities, conservation costs, a phylogenetic tree) will not be known with any degree of certainty. In chapter 8 we investigate the effect of using incorrect species survival probabilities. We show that solutions are robust to uncertainty in these parameters, as long as non-zero survival probabilities are specified. We develop an extension to the NAP that permits the uncertainty in the survival probabilities to be explicitly included. For several examples we show that simple point estimates for the probabilities produced solutions that are nearly as good as those produced using our extension of the NAP. This is beneficial for practical conservation management problems as the simpler conceptual approach can be used without too great a penalty.

Incorporating a formal measure of biodiversity in a conservation management framework is a complex task. An important question therefore, is whether the gain this achieves makes it a worthwhile exercise? In chapter 9 we introduce two upper bounds on the expected achievable gain of using PD to prioritise species for conservation. The typical increase in the conserved biodiversity (for both Yule trees and a sample of 'real' trees) was about 20% however gains as high as 150% were achievable for some examples. Lastly we note that potential gains are highest for imbalanced trees, as it is crucial to identify the 'oldest' species.

Chapter VI

Phylogenetic diversity

Phylogenetic Diversity (PD; Faith (1992)) is a measure of the biodiversity of a set of species. PD has been used in a wide variety of applications including biodiversity conservation (eg. Crozier et al. (2005), Lewis and Lewis (2005), Mooers et al. (2005), Soutullo et al. (2005), and Faith and Williams (2006)) and prioritizing taxa for genomic sequencing (Pardi and Goldman, 2005). PD is calculated from the phylogenetic tree, $\mathcal{T}$, the leaves of which correspond to the set of taxa, X, of interest. For a subset, Y of X the PD is the sum of the branch lengths of the phylogenetic tree containing taxa in Y and the root, an example is given in Figure 6.1.

Formally, the phylogenetic diversity of Y is denoted by $PD(Y)$ and denoting the length of an edge e of $\mathcal{T}$ by λ_e we have:

$$PD(Y) = \sum_e \lambda_e,$$

where the summation is over all edges e in $\mathcal{T}$ that lie on the minimal subtree of $\mathcal{T}$ connecting the taxa in Y (and if $\mathcal{T}$ is rooted, also connecting the root). There has been some debate about whether the root should be included, however the original definition in Faith (1992) and prevailing usage include the root (see Crozier et al. (2005), Faith and Baker (2006) and Crozier et al. (2006) for further discussion).

Depending on the data from which a tree is derived, the branch lengths may have different interpretations. Branch lengths may correspond to an evolutionary time-scale (i.e. the number of millions of years between speciation events), or to genetic distance, or to the extent of morphological differences, or perhaps some combination of these (or other) measures of evolutionary distance. Throughout this thesis, no particular interpretation is assumed, so

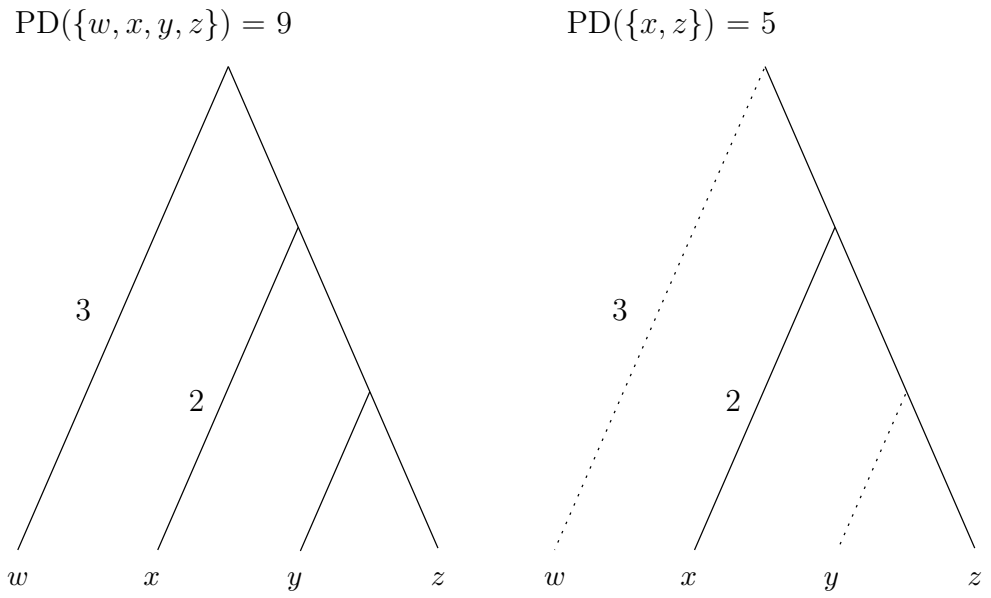


Figure 6.1: Left: The PD of the set of all species in this tree is found by adding all edge lengths (which are length one unless otherwise indicated). The PD of a subset of species is the sum of all edges spanned by the subset and the root. Right: For a subset $Y = \{x, z\}$, $PD(Y)$ is found by adding the lengths of the solid lines of the tree on the right.

as to allow the greatest degree of generality for applications; in particular, unless we state so explicitly, we do not assume that the tree is ultrametric (an *ultrametric tree* is one for which the distance from the root to any leaf is the same, as would occur for (a) genetic distance under a 'molecular clock', or (b) an evolutionary time-scale).

The PD measure has several combinatorial and algorithmic properties which we now describe.

6.1 Combinatorial and algorithmic properties

6.1.1 Generalized Pauplin formula

PD can be written canonically as a linear combination of pairwise distances within the tree. That is, if $d(x, y)$ denotes the distance between x and y in $\mathcal{T}$, the PD of a set W can be written as

$$PD(W) = \sum_{x,y \subseteq W} \mu_{\mathcal{T},W}(x, y)d(x, y) \tag{6.1}$$

where $\mu_{\mathcal{T},W}$ is a function that depends on $\mathcal{T}$ and W but not the branch lengths. Actually there are many possible choices of $\mu_{\mathcal{T},W}$ but there is one that is particularly natural and which is defined as follows. Let $\mathcal{T}_W$ denote the subtree of $\mathcal{T}$ connecting W and let $p(\mathcal{T}_W, x, y)$ be the set of non-leaf vertices of $\mathcal{T}_W$ that lie on the path connecting x and y. Then set

$$\mu_{\mathcal{T},W}(x, y) = \prod_{v \in p(\mathcal{T}_W, x, y)} (d(v) - 1)^{-1}$$

where $d(v)$ is the degree of vertex v in $\mathcal{T}_W$. The validity of Equation 6.1 for this choice of $\mu_{\mathcal{T},W}$ was described (for $W = X$) for binary phylogenetic X–trees in Pauplin (2000), and generalized to arbitrary phylogenetic X–trees in Semple and Steel (2004). The Pauplin formula also provides an interesting starting point for forming species specific indices of biodiversity such as the Equal-Splits index (chapter 10).

6.1.2 The strong exchange property

For any function f defined from the collection of subsets of X of size at least r into the real numbers, we say that f satisfies the *strong exchange property* if for any two subsets Y and Z with $r \leq |Y| < |Z|$ there exists some taxon $z \in Z - Y$ such that:

$$f(Z - \{z\}) - f(Z) + f(Y \cup \{z\}) - f(Y) \geq 0. \tag{6.2}$$

The strong exchange property was established for $f = PD$ (and $r = 2$ in the case of unrooted phylogenetic trees) in Steel (2005); its interpretation in this setting is that for any two of the subsets, the larger one contains some taxon (z) that would contribute at least as much to the PD value of the smaller subset than it adds to that of the larger one.

Consider both trees in Figure 6.1 and the situation where the two subsets Y and Z are $\{z\}$ and $\{w, y\}$ respectively; clearly $|Y| < |Z|$. Deleting taxon y from subset Z and adding it to Y results in a loss of the combined PD of Y and Z in both trees, hence y does not satisfy the strong exchange property. However the combined PD of Y and Z remains the same if taxon w is removed from Z and added to Y, thus satisfying the strong exchange property. Note that the strong exchange property for PD fails for $r = 1$ and $r = 0$ for unrooted trees, but holds for rooted trees.

6.1.3 Finding sets of maximal PD

PD is a measure of biodiversity, hence finding sets of species that maximise PD is an interesting problem from a biodiversity conservation perspective. All else being equal, those species with the highest PD should be conserved. Steel (2005) established the previous strong exchange property for PD and showed that this property is a sufficient condition for a greedy algorithm to produce sets of species with maximal PD. This follows by standard arguments from 'greedoid' theory (see Korte et al. (1991)). To construct such a subset the greedy algorithm iteratively adds the element (taxon) that gives a maximal increase in f until the subset contains r elements. Formally from Steel (2005) for unrooted trees we have:

Theorem 7. *Each PD maximising set, S, of size k ($k \geq 2$) can be constructed in the following way. Begin with S containing two of the species that are the farthest apart in the tree. Sequentially add one of the taxa to S that gives the greatest increase in the PD score. Continue adding taxa in this way until k elements have been selected.*

Note that all PD maximising sets are obtainable by making different arbitrary choices when several species could have been added to S. This theorem can also be applied to rooted trees and the unconventional definition of PD that does not automatically include the root. For rooted trees and the more usual definition of PD (which always includes the root) the greedy algorithm is readily adapted:

Theorem 8. *Each PD maximising set, S, of size k ($k \geq 2$) can be constructed in the following way. Begin with an empty set S. Sequentially add one of the taxa to S that gives the greatest increase in the PD score. This is repeated until S contains k species.*

Moreover, as demonstrated in Pardi and Goldman (2005), for any given set of taxa W of size at least 2 (or 1 in case of rooted trees) the strong exchange property also ensures that amongst the collection of all subsets of size k containing W, the one(s) of maximal PD value can be constructed from W by the greedy algorithm (even though W itself may not have optimal PD score for its cardinality). In a conservation setting this means that if some species have already been selected (eg. for their charismatic value or ecological importance) then the best complementing set of species can be obtained using the greedy algorithm. This is also useful in a genomic sequencing setting where some sequences will have already been obtained and we wish to prioritise species for future sequencing (Pardi and Goldman, 2005).

6.1.4 Finding sets of minimal PD

Sets of species with minimal PD correspond to the worst possible choice for conservation. As such they provide a baseline with which to compare the benefit provided by other solutions. For example if an ultrametric tree is

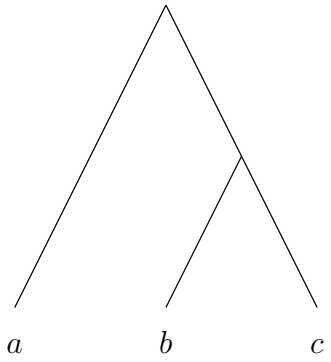


Figure 6.2: This is the simplest example where the optimal substructure is violated for PD minimisation. An optimal subset of size one is $\{a\}$. The optimal substructure property therefore requires the existence of a set of size two that contains $\{a\}$. However the only optimal set of size two is $\{b, c\}$ hence the optimal substructure property is violated and the greedy algorithm is not guaranteed to produce optimal solutions (the sub-optimal sets $\{a, b\}$ and $\{a, c\}$ can be obtained by the greedy algorithm).

star-like, for a given set size the difference between the PD minimising and maximising solution will be minimal, hence the benefit of using advanced methods to select species is negligible. This application will be considered in more detail in chapter 9, here we introduce a simple algorithm for finding PD minimising sets.

Firstly we note that the strong exchange property is not satisfied in this situation. Minimising PD is equivalent to maximising negative PD; it is easy to construct an example where the strong exchange property is not satisfied for negative PD. A further property that is required for a greedy algorithm to be able to produce optimal solutions is the optimal substructure property. The optimal substructure property states that an optimal set of size k must contain optimal subsets of all sizes less than k and must itself be contained within an optimal subset of all sizes larger than k. Figure 6.2 shows an example where the optimal substructure property is not satisfied for the PD minimising problem and a greedy algorithm therefore does not produce optimal PD minimising sets.

We introduce a simple dynamic programming algorithm that can be used to produce PD minimising sets. The algorithm works from the leaves upwards. At each internal node, i, the optimal subsets from the two subtrees are combined and compared to find the PD minimising sets of size one through C_i. We denote the PD minimising set of size k for an internal node i by $S_{\min}(i, k)$. Formally the dynamic programming algorithm is implemented by conducting a depth first traversal over the internal nodes in $\mathcal{T}$ and applying the following steps at each internal node, i:

For each k from 0 to C_i:

- Find a j that minimises $\hat{PD}_i(S_{\min}(a,j) \cup S_{\min}(b,k-j))$.
- Using that j set $S_{\min}(i,k) = S_{\min}(a,j) \cup S_{\min}(b,k-j)$.

Where $\hat{PD}_i$ is the PD calculated on the tree subtended by i (including its root edge). For the leaves $S_{\min}(i,1)$ is trivially $\{i\}$ and by definition $\hat{PD}_i(\{i\})$ is equal to the pendant edge of i. Upon completion this will yield the PD minimising sets of all subset sizes at the root node. The algorithm is illustrated in Figure 6.3.

6.1.5 Exclusive molecular phylodiversity

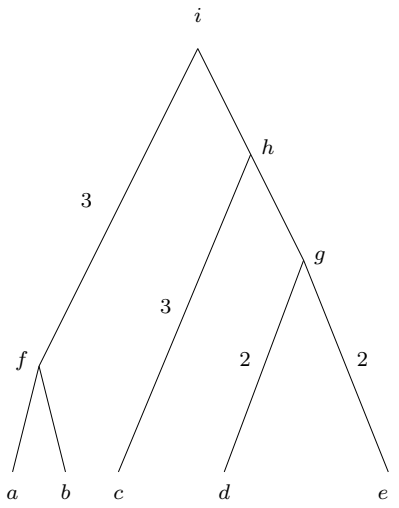
Lewis and Lewis (2005) recently investigated a measured related to PD which they called the 'exclusive molecular phylodiversity' (EPD) of a set Y, defined by:

$$EPD(Y) := PD(X) - PD(X - Y).$$

This measure has also been used by Sechrest et al. (2002) to assess the evolutionary history of endemic species in biodiversity hotspots. The benefit of exclusive molecular phylodiversity in that context is that it avoids the need for any information about non-endemic species, effectively assuming that these are well represented elsewhere. It is easy to show that this measure does not satisfy the optimal substructure property and therefore greedy algorithms cannot be guaranteed to produce an optimal subset, Y, (see Figure 6.4 for an example where the greedy algorithm does not work).

6.2 Loss of phylogenetic diversity under extinction models

Nee and May (1997) investigated the loss of PD as taxa are randomly deleted from random trees under a simple model: each taxon is equally likely to be the next to become extinct (the 'field of bullets' model). The trees were ultrametric trees as generated by a random birth model. They found a characteristic concave shape in the relationship between the expected remaining PD and the proportion of taxa deleted. This relationship is illustrated for the Crested Penguins tree (Figure 6.5) by the upper curve in Figure 6.6.



$\hat{PD}(S_{\min}(i,k))$

Node (i) \ Set size (k)	1	2	3	4	5
a	1	-	-	-	-
b	1	-	-	-	-
c	3	-	-	-	-
d	2	-	-	-	-
e	2	-	-	-	-
f	4	5	-	-	-
g	3	5	-	-	-
h	4	6	9	-	-
i	4	5	9	11	14

$S_{\min}(i,k)$

Node (i) \ Set size (k)	1	2	3	4	5
a	$\{a\}$	-	-	-	-
b	$\{b\}$	-	-	-	-
c	$\{c\}$	-	-	-	-
d	$\{d\}$	-	-	-	-
e	$\{e\}$	-	-	-	-
f	$\{a\}$	$\{a,b\}$	-	-	-
g	$\{d\}$	$\{d,e\}$	-	-	-
h	$\{c\}$	$\{d,e\}$	$\{c,d,e\}$	-	-
i	$\{a\}$	$\{a,b\}$	$\{c,d,e\}$	$\{a,b,d,e\}$	$\{a,b,c,d,e\}$

Figure 6.3: This figure gives a simple example of the PD minimising algorithm. Unlabelled edges in the tree have length one and the vertices are labelled a through i. The bottom panel shows the PD minimising sets, where multiple sets were optimal one was chosen by lexical order.
Consider node i. An optimal set of size 2 is found by considering the optimal sets of nodes h and f that when combined contain 2 species. The possibilities are to select 0, 1 or 2 species from h and respectively 2, 1 or 0 species from f. The lowest PD is obtained by selecting two species from f (and none from n), consequently the optimal subset of size 2 is $\{a,b\}$.

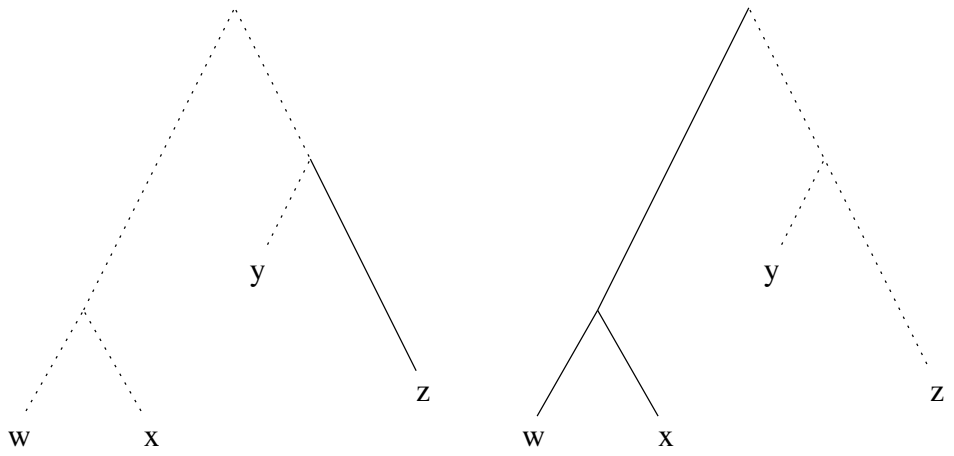


Figure 6.4: Left: The single species with maximal EPD is z, as it has the longest pendant edge. Right: The subset of two species with maximal EPD is $\{w, x\}$ as the EPD of all species in a clade (in this case a cherry) also includes the edge connecting that clade to the rest of the tree (in this case the root). In this case the optimal substructure property is not satisfied and greedy algorithms are not guaranteed to produce sets of species with optimal EPD.

This relationship was further investigated recently in Soutullo et al. (2005), which studied random deletion of taxa from certain biological trees. Once again the relationship between taxa deleted and remaining PD was concave. Recall that a sequence $x = (x_1, x_2, \ldots, x_n)$ of real numbers is *concave* if, when we let $\Delta x_r = x_r - x_{r-1}$ the following inequality holds for all r:

$$\Delta x_r - \Delta x_{r+1} \geq 0$$

and the sequence is *strictly concave* if the inequality is strict for all r. Geometrically this means that the slope of the line joining adjacent points in the graph of x_r versus r is decreasing. Note that x_r is concave precisely if the complementary (reverse) sequence $y_r = x_{n-r}$ is concave. The significance of (strict) concavity for PD is that it says (informally) that most PD loss comes near the end of an extinction process.

In this section we first describe a generic concave relationship observed between the average PD and the number of taxa deleted. This makes intuitive sense, because each interior branch survives until the point where there

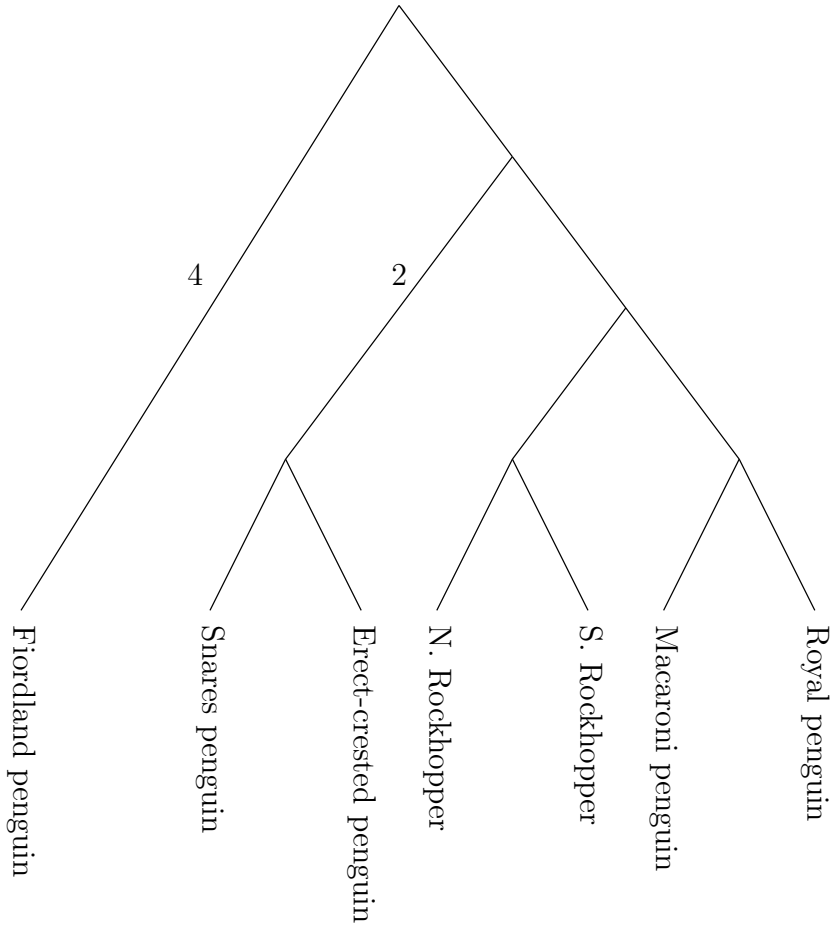


Figure 6.5: The phylogenetic tree for crested penguins. This tree was derived from the tree in Bertelli and Giannini (2005) and Giannini and Bertelli (2004) which had no branch lengths. For illustrative purposes each level in the original tree was assumed to be separated by the same distance such that all edges in this tree are of length 1 except for the two marked edges.

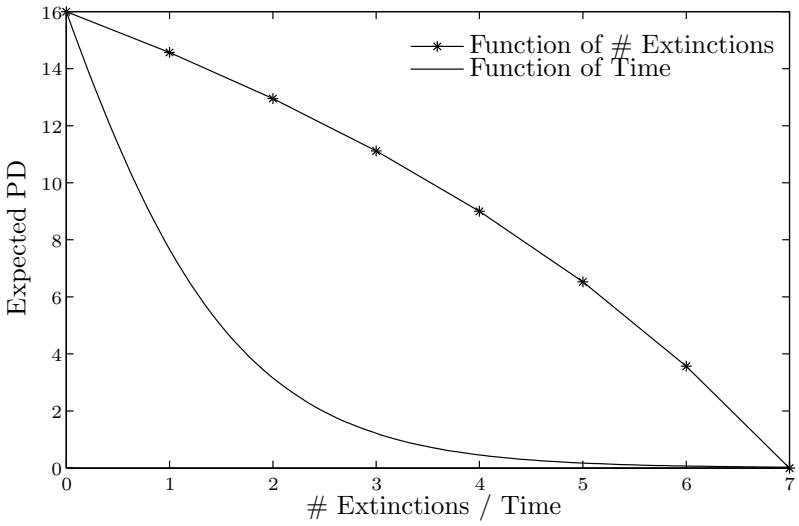


Figure 6.6: The expected remaining PD after extinctions have occurred among the crested penguins depicted in Figure 6.5. This loss in PD is viewed as a function of both the number of extinctions that have occurred and the time that has elapsed since extinctions began occurring.

is no taxon below it and this is likely to occur towards the end of a random extinction process.

Consider a rooted phylogenetic tree having a leaf set X of size n. Let W be a random subset of taxa of size r sampled uniformly from X (for example, by selecting uniformly at random a set S of $n - r \geq 0$ elements of X and deleting them, in which case $W = X - S$). For $r \in \{1, \ldots, n\}$ let $\mu_r = \mathbb{E}[PD|r]$, the expected value of $PD(W)$ over all such choices of W. Equivalently, we can write $\mu_r = \binom{n}{r}^{-1} \sum_{W \subseteq X : |W|=r} PD(W)$, where $\binom{n}{r}$ is the binomial coefficient ($= \frac{n!}{r!(n-r)!}$), which is the number of ways of selecting r elements from a set of size n. For brevity we adopt the usual convention that $\binom{n}{r} = 0$ if r is greater than n or less than 0.

Clearly $\mu_n = PD(X)$. For $r \in \{1, \ldots, n\}$, let $\Delta\mu_r = \mu_r - \mu_{r-1}$. Note that, since $\mu_0 = 0$, we have $\Delta\mu_1 = \mu_1$. For an edge e of $\mathcal{T}$, and $r \in \{1, \ldots, n-1\}$ let

$$\psi(e, r) := \frac{C_e(C_e - 1)}{r(r+1)} \cdot \frac{\binom{n-C_e}{r-1}}{\binom{n}{r+1}}$$

where C_e denotes the number of leaves of $\mathcal{T}$ that lie 'below' e (i.e. separated from the root by e).

The following theorem is reproduced here from Steel (2006). It shows that for any fully resolved tree PD decays in a strictly concave fashion as taxa are randomly deleted, and the only trees for which the decay of PD is linear are fully unresolved 'star' trees. In the following theorem a *cherry* is a pair of leaves that are adjacent to the same vertex.

Theorem 9. *Consider a phylogenetic tree $\mathcal{T}$ with an assignment λ of positive branch lengths. Then, for each $r \in \{1, \ldots, n-1\}$,*

$$\Delta\mu_r - \Delta\mu_{r+1} = \sum_e \lambda_e \psi(e, r)$$

where the summation is over all edges of $\mathcal{T}$. In particular, μ is concave over this domain, and μ is strictly concave if and only if $\mathcal{T}$ has a cherry, while μ is linear if and only if $\mathcal{T}$ has no interior edges (i.e. is an unresolved 'star' tree).

Consider the tree for crested penguins to which we have previously referred (Figure 6.5). Figure 6.6 shows the expected PD as a function of the number of extinctions. As expected from the above theorem, the relationship depicted in this figure is strictly concave.

6.2.1 Relationship between PD and time under extinction

We have investigated the expected PD as a function of the number of extinctions that have occurred. So far each taxon has been considered as equally likely to be the next to become extinct. However no consideration has been given to the timing of these extinctions. Here we consider the situation where each taxon has the same probability of becoming extinct at any point in time (the time to extinction for an individual taxon has an exponential distribution) and consider the expected PD as a function of the time instead of the number of extinctions that have occurred. We will show that the decline in expected PD does not in general have a concave shape and in fact after a specific time (dependent on the tree shape) the decline will become convex. Note that this is not a contradiction with the previous result; it is simply

due to the fact that the number of extinctions decreases over time as there are fewer species left that could become extinct.

The probability that an edge, e, will be spanned by the taxa remaining at some time t, depends only on the number of children (C_e) of that edge. Denoting this probability by $p_e(t)$ we have:

$$p_e(t) = 1 - \left(1 - e^{-rt}\right)^{C_e}$$

where r is the rate of extinction. The expected PD at time t, $\mathbb{E}_t(PD)$ is easily found using these probabilities:

$$\mathbb{E}_t(PD) = \sum_e \lambda_e p_e(t).$$

Observe that $\mathbb{E}_t(PD)$ depends only on the sums of the edges with the same number of leaves attached, not on the individual edges themselves:

$$\mathbb{E}_t(PD) = \sum_{j=1}^{m} \alpha_j \left[1 - \left(1 - e^{-rt}\right)^j\right],$$

where $\alpha_j = \sum_{e, C_e = j} \lambda_e$, and m is the highest number of leaves below any edge – this corresponds to the edge(s) at the root with the most leaves descendant from them. To investigate the shape of $\mathbb{E}_t(PD)$ the second derivative is easily obtained:

$$\frac{d^2\mathbb{E}_t(PD)}{dt^2} = r^2 e^{-rt}\left(\alpha_1 + \sum_{j=2}^{m} \alpha_j j \left(1 - je^{-rt}\right)\left(1 - e^{-rt}\right)^{j-2}\right). \quad (6.3)$$

For convexity, the second derivative must be positive. The term corresponding to α_1 is clearly positive, but the sign corresponding to the other α-values depends on t. The term corresponding to a particular α_j is positive if $1 - je^{-rt} > 0$ which holds when

$$t > \frac{\ln(j)}{r}.$$

A sum of convex functions is convex, therefore once the above condition is

satisfied for all j, $\mathbb{E}_t(PD)$ will be convex. The term that becomes convex the latest is the term with the highest value of j (namely m). Convexity is therefore guaranteed after $\hat{t} = \ln(m)/r$. In the limit as $\sum_{j<m} \alpha_j/\alpha_m \to 0$, $PD(t)$ will become convex exactly at $\hat{t}$, however $PD(t)$ will generally become convex earlier due to the other terms.

The terms corresponding to edges with high values of j are the last to become positive; as more weight is assigned to these the time to convexity lengthens. Variation in diversification rates through time and/or among clades can therefore affect the time to convexity.

To obtain the exact time to convexity is non-trivial as it involves calculating the highest root of Equation 6.3 - a polynomial of order $m-1$ in e^{-rt}. Common techniques for bounding the roots (eg. the Cauchy Bound) can be applied by transforming the polynomial appropriately. No general properties were derived from these techniques as they depend strongly on the values of the α_j's.

The amount of PD loss that has occurred by the time that convexity is guaranteed ($\hat{t} = \ln(m)/r$) is difficult to characterize, but the number of taxa remaining at this time can readily be found. The probability of an individual taxon persisting to time t is e^{-rt}, so at $t = \hat{t}$ each taxon is extant with probability $1/m$. The total number of taxa is between $m+1$ and $2m$ (depending on the imbalance of the tree at the root) and the expected number of extant taxa at $t = \hat{t}$ is therefore between 1 and 2. Accordingly the convexity result may appear to be of limited biological interest, however given a real tree the expected number of taxa remaining by the time convexity is reached will usually be much higher.

Another interesting behaviour that can readily be examined and may be of more practical interest is the initial shape of the PD decline (that is at and just after $t = 0$). Substituting $t = 0$ in Equation 6.3 we obtain:

$$\begin{aligned} \frac{d^2\mathbb{E}_t(PD)}{dt^2}|_{t=0} &= r^2 \left(\alpha_1 + \sum_{j=2}^{m} \left(\alpha_j j \left(1-j\right) 0^{j-2} \right) \right) \\ &= r^2 \left(\alpha_1 - 2\alpha_2 \right). \end{aligned} \tag{6.4}$$

Initial convexity requires $\alpha_1 > 2\alpha_2$ and concavity requires $\alpha_1 < 2\alpha_2$. The edges that contribute to α_1 are the pendant edges and those contributing to α_2 are edges above cherries. Any tree can have at most half as many 'above cherry' edges as pendant edges, so if pendant edges have similar lengths as the 'above cherry' edges then that tree will therefore exhibit initial convexity (as for the Crested Penguins tree Figures 6.5 and 6.6). It should be noted that even if the PD loss curve for a tree is convex at $t = 0$ and after $t = \hat{t}$ there is no guarantee that it will be convex between these two times due to the complexity of Equation 6.3.

PD loss for Yule trees

The time to guaranteed convexity, $\hat{t}$, is an overestimate of the true time as this result applies to all possible phylogenetic tree shapes and edge lengths. For most trees it is expected that convexity will be reached much earlier. It is therefore of interest to explore the second derivative of the PD loss function for trees produced by the Yule model since these give some approximation of the trees expected to be found in nature. Using either of the methods developed in chapter 2 we can easily derive the expectation of Equation 6.3 for trees produced by a Yule model. This is straightforward due to the linearity (in α_j) of Equation 6.3, the result is shown in Figure 6.7A. This indicates that for most tree shapes the expected PD loss will be convex for all times t. It should be noted that for any tree shape we can choose edge lengths (which may be very improbably under the Yule model) such that the PD loss will be concave at some time.

Whenever two species/leaves are directly descendant from the same ancestor we refer to them as a cherry (McKenzie and Steel, 2000). As discussed in Hartmann and Steel (2007) the number of cherries contributes to initial convexity and Figure 6.7B shows that this also holds for Yule trees in our situation. Unless a ten species tree has at least four cherries it is expected to exhibit initial convexity. Approximately 20% of observed tree shapes are expected to exhibit initial concavity.

The extinction process considered here is very simple – all species are at the same risk of becoming extinct throughout time. In reality it is expected

that these risks will not be independent, for example if a species becomes extinct other species dependent on it will have an increased risk of extinction or its competitors may have a decreased risk of extinction. This may also have a temporal effect on the extinction process, after many extinction events have occurred the extinction risk for the remaining species may increase due to their interdependency. The effect of interdependent extinction risks on the loss of PD is worthy of further exploration for which our methods may help yield greater insight.

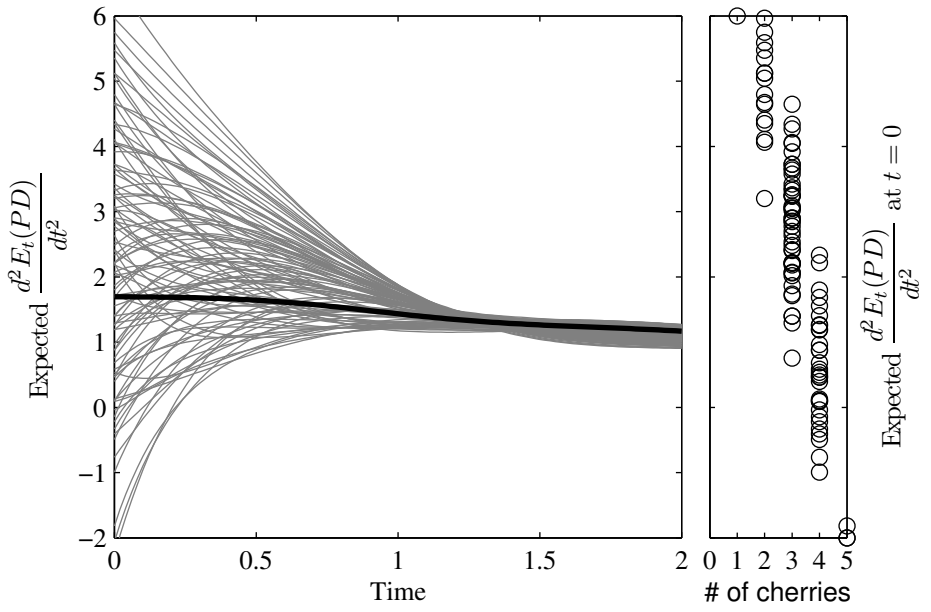


Figure 6.7: The expected second derivative of the PD loss curve is shown here for trees with ten species generated by the Yule model. The gray lines on the left panel are conditional on individual tree shapes and the solid line is the weighted mean over all tree shapes. Only 20% of tree shapes are expected to have a negative derivative at some time, the remainder are expected to be convex for all times. The right panel shows the number of cherries for each tree shape as a function of the initial expected second derivative for that tree shape (as shown on the left panel). As suggested in Hartmann and Steel (2007) this indicates that a large number of cherries is required for initial concavity. The trees are conditioned on having the most likely age for their size ($t = \log(10)$) and the time to extinction is exponentially distributed with rate 1.

Chapter VII

The Noah's Ark Problem

Biodiversity conservation requires a methodology for prioritizing the taxa to conserve, given limited resources. Many conservation approaches have simply aimed to conserve as many taxa as possible (Gaston, 1996), however a more appropriate method should take taxon distinctiveness into account (for review, see Crozier (1997)) and aim to minimize the future loss of biodiversity. Witting and Loeschcke (1995) (see also Witting et al. (2000)) linked the phylogenetic diversity (PD; chapter 6, Faith (1992)) measure to extinction probabilities to obtain a method for minimizing the future loss of biodiversity.

Consider a situation where each taxon, j, has some probability, a_j, of remaining extant until some given future time. To compare different conservation approaches (and their corresponding species survival probabilities) we need to find the expected PD of the surviving species. A particular branch length is included in the final PD score if at least one of the children of that edge remains extant. For example the edge connecting y and z with the rest of the tree in Figure 6.1 will be preserved as long as one of its children (taxa y or z) remains extant. If these taxa both have a survival probability of 0.9, the probability that at least one will remain extant is simply $1-(1-0.9)^2 = 0.99$. Denoting the children of a particular edge, i, by C_i the expected PD can be expressed as:

$$\mathbb{E}(PD) := \sum_i \lambda_i \left(1 - \prod_{j \in \mathcal{C}_i} (1 - a_j)\right), \tag{7.1}$$

where λ_i is the length of edge i, and the summation is over all edges of $\mathcal{T}$.

Weitzman (1998) proposed the "Noah's Ark Problem" (NAP), a framework based on PD that incorporates costs and probabilities, and has seen some practical application including conservation of cattle breeds

(Simianer et al., 2003; Reist-Marti et al., 2006). In the NAP each taxon has a survival probability which can be increased at some cost. The objective is to allocate a limited budget to the taxa such that the future expected biodiversity – $\mathbb{E}(PD)$ is maximized.

Weitzman (1998) showed that the optimal solution will be extreme – each species is either fully conserved or not at all. Consequently the problem is simplified from one of finding an optimal budgetary allocation to one of finding an optimal set of species to conserve. Despite this simplification obtaining optimal resource allocations is still a complex problem and it may be necessary to consider a large proportion of the possible subsets of the N taxa that can be conserved. The number of such subsets grows at rate 2^N, consequently for problems involving more than a few dozen taxa it is not computationally feasible to consider all subsets and an efficient algorithm is required for obtaining optimal solutions to the NAP.

Suggestions have been made in the literature that any NAP for which the associated tree satisfies a molecular clock can be solved using a greedy algorithm (like that introduced in the previous chapter) (Simianer et al., 2003). Several aspects of the NAP which prevent the greedy algorithm from producing optimal solutions in all cases are examined here. These examples (Figures 7.1 and 7.4) illustrate that a greedy algorithm is not, in general, guaranteed to produce optimal solutions. However we show that greedy algorithms can produce optimal solutions to the two restricted variations of the NAP and one extension to it.

7.1 Formal definition

A variation of the Noah's Ark Problem was described in Weitzman (1992), however this used a measure of dissimilarity instead of PD (see Faith et al. (2003) for a discussion); the NAP as published in Weitzman (1998) finally combined PD, extinction probabilities and conservation costs.

In the NAP framework each taxon, j, has some probability, a_j, of remaining extant, however if some conservation intervention of cost c_j is applied to this taxon, then this survival probability can be increased from a_j to b_j. Given a budgetary constraint, B, the problem is to find the set of taxa to

conserve, S, that maximizes the future expected phylogenetic diversity, denoted by $\mathbb{E}(PD|S)$. $\mathbb{E}(PD|S)$ is calculated by summing all the edge lengths, λ_i, in the tree, weighting each edge by the probability that it will be spanned by the surviving taxa:

$$\begin{aligned}\mathbb{E}(PD|S) &:= \sum_i \lambda_i p(i|S) \\ &= \sum_i \lambda_i \left(1 - \prod_{k \in \mathcal{C}_i - S} (1 - a_k) \prod_{l \in \mathcal{C}_i \cap S} (1 - b_l)\right), \qquad (7.2)\end{aligned}$$

where $p(i|S)$ denotes the probability that one of the taxa in $\mathcal{C}_i$ will remain extant given that the set of taxa S is being conserved.

The formulation of the NAP used throughout this thesis is essentially equivalent to that given in Weitzman (1998) but is expressed differently for convenience:

Given an edge-weighted phylogenetic tree, and values (a_j, b_j, c_j) for each taxon j, maximize $\mathbb{E}(PD|S)$ over all subsets S of taxa, subject to the constraint: $\sum_{j \in S} c_j \leq B$.

The constraint ensures that the cost of conserving the taxa in S does not exceed the budget (B).

The original formulation of the NAP included an additional term in the objective function that permitted each taxon to have an intrinsic value (utility) unrelated to its contribution to PD (eg. the value of tourism for a species of whale). This additional value has not been made explicit here as it is easy to show that including such a value for a particular taxon is equivalent to adding it to the length of the pendant edge for that taxon.

The original formulation of the NAP also allowed taxa to be partially protected, so that resources could be spread more thinly across multiple taxa instead of conserving a smaller subset of taxa to the maximum extent possible. Weitzman (1998) assumed that if a taxon is partially conserved the survival probability increase for that taxon is directly proportional to the proportion of the funding that taxon received. If q_j is spent on conserving

taxon j ($0 \leq q_j \leq c_j$) the new survival probability, $g_j(q_j)$, for taxon j is:

$$g_j(q_j) = \frac{q_j}{c_j}(b_j - a_j) + a_j. \tag{7.3}$$

Weitzman (1998) showed that under this assumption, the solutions to the NAP are extreme – the optimal solution will always allocate the maximum amount ($q_j = c_j$) to a few taxa instead of partially conserving ($0 < q_j < c_j$) a greater number of taxa, with the possible exception of the last taxon conserved which may only be partially conserved due to budgetary constraints. Consequently the problem of deciding how much funding to allocate to the conservation of each taxon becomes a problem of deciding which taxa to conserve. Throughout this chapter we adopt the convention that the last taxon selected for conservation will be partially conserved such that the full conservation budget is utilized.

The benefit of Equation 7.3 is further demonstrable by considering a star tree (each taxon is directly descendant from the root) where the taxa may have different costs. If $a_j = 0$ and $b_j = 1$ for all taxa j and each taxon can be either fully conserved or not at all:

$$g_j(q_j) = \begin{cases} 0, & q_j < c_j; \\ 1, & q_j \geq c_j; \end{cases}$$

then the problem is equivalent to the "knapsack problem" which is well known to be NP-complete (Cormen et al., 2002). However if $g_j(q_j)$ is given by Equation 7.3 instead, the problem is equivalent to the "fractional knapsack" problem which is solvable by a greedy algorithm (Cormen et al., 2002).

7.1.1 Extremality of solutions to the generalised NAP

We refer to problems where the survival probability relationships do not satisfy Equation 7.3 as generalised Noah's Ark Problems (g-NAPs). In this situation we denote the probability that a species, i, survives given that the expenditure on all species is $\vec{q}$ by $g_i(\vec{q})$, note that $g_i(\vec{q})$ is only defined on the domain of possible expenditures ($0 \leq q_j \leq c_j, \forall j$). For convenience we also adopt the notation that $\frac{\partial g_i(\vec{q})}{\partial q_j} = g_{i,j}$.

In terms of these survival probabilities the expected PD over a set of trees, T, for a given expenditure $\vec{q}$ is:

$$\mathbb{E}(PD) = \sum_{\mathcal{T} \in T} p_{\mathcal{T}} \sum_i \lambda_i \left[1 - \prod_{j \in \mathcal{C}_i} (1 - g_j(\vec{q})) \right], \tag{7.4}$$

where $p_{\mathcal{T}}$ is the probability associated with tree $\mathcal{T}$. The objective in this generalised NAP setting is, as before, to maximise $\mathbb{E}(PD)$ subject to the sum of the conservation costs ($\sum_i g_i(\vec{q})$) not exceeding a budget, B.

Theorem 10. *The set of optimal solutions to the above problem contains an extreme solution if for all $\alpha, \beta, j, n \in X$:*

- *$g_j(\vec{q})$ is convex*
- *$g_{j,\alpha} g_{n,\beta} + g_{j,\beta} g_{n,\alpha} - g_{j,\alpha} g_{n\alpha} - g_{j,\beta} g_{n,\beta} \geq 0$*

These conditions are trivially satisfied if the species survival probabilities satisfy Equation 7.3 (the conventional NAP).

Proof. The proof works by showing that given any non-extreme budget allocation, $\vec{q}$, there exists an extreme solution that is at least equally as good.

Consider an optimal non-extreme budget allocation $\hat{\vec{q}}$ and let α and β be two of the species that are partially conserved (there may or may not be others). We want to show that by spending less on one species and more on the other species we will obtain an increase in $\mathbb{E}(PD)$. Making the dependence of $\mathbb{E}(PD)$ on q_α and q_β explicit and shifting a budget of δ from species β to α we can obtain the following Taylor series expansion:

$$\begin{aligned}\mathbb{E}(PD|q_\alpha + \delta, q_\beta - \delta) = \mathbb{E}(PD|q_\alpha, q_\beta) + \delta\Big(&\mathbb{E}(PD|q_\alpha, q_\beta)_\alpha - \\ &\mathbb{E}(PD|q_\alpha, q_\beta)_\beta\Big) + \frac{\delta^2}{2}\Big(\mathbb{E}(PD|\acute{q_\alpha}, \acute{q_\beta})_{\alpha\alpha} + \mathbb{E}(PD|\acute{q_\alpha}, \acute{q_\beta})_{\beta\beta} - \\ &2\mathbb{E}(PD|\acute{q_\alpha}, \acute{q_\beta})_{\alpha\beta}\Big)\end{aligned}$$

where $q_\alpha \leq \acute{q_\alpha} \leq q_\alpha + \delta$ and $q_\beta - \delta \leq \acute{q_\beta} \leq q_\beta$ for positive δ and similarly for negative δ. We are free to choose the sign of δ such that the term involving

the first derivative is positive, hence, for this choice of δ, $\mathbb{E}(PD|q_\alpha+\delta, q_\beta-\delta)$ will be greater than $\mathbb{E}(PD|q_\alpha, q_\beta)$ if:

$$f := \frac{\delta^2}{2}\Big(\mathbb{E}(PD|q_\alpha', q_\beta')_{q_\alpha q_\alpha} + \mathbb{E}(PD|q_\alpha', q_\beta')_{q_\beta q_\beta} - 2\mathbb{E}(PD|q_\alpha', q_\beta')_{q_\alpha q_\beta}\Big) \geq 0.$$

The relevant derivatives are:

$$\begin{aligned}
\mathbb{E}(PD|q_\alpha, q_\beta)_{q_\alpha q_\alpha} &= \sum_{\mathcal{T}\in T}\sum_i \lambda_i p_{\mathcal{T}}\Bigg(\sum_{j\in\mathcal{C}_i} g_{j,\alpha\alpha}\prod_{m\in\mathcal{C}_i-\{j\}}[1-g_m] - \\
&\qquad 2\sum_{j,n\in\mathcal{C}_i} g_{j,\alpha}g_{n,\alpha}\prod_{m\in\mathcal{C}_i-\{j,n\}}[1-g_m]\Bigg) \\
\mathbb{E}(PD|q_\alpha, q_\beta)_{q_\alpha q_\beta} &= \sum_{\mathcal{T}\in T}\sum_i \lambda_i p_{\mathcal{T}}\Bigg(\sum_{j\in\mathcal{C}_i} g^j_{\alpha\beta}\prod_{m\in\mathcal{C}_i-\{j\}}[1-g_m] - \\
&\qquad \sum_{j,n\in\mathcal{C}_i}(g_{j,\alpha}g_{n,\beta}+g_{j,\beta}g_{n,\alpha})\prod_{m\in\mathcal{C}_i-\{j,n\}}[1-g_m]\Bigg).
\end{aligned}$$

These derivatives can be substituted in f and after some basic manipulation we obtain:

$$\begin{aligned}
f &= \frac{\delta^2}{2}\sum_{\mathcal{T}\in T}\sum_i \lambda_i p_{\mathcal{T}}\Bigg(\sum_{j\in\mathcal{C}_i}(g_{j,\alpha\alpha}+g_{j,\beta\beta}-2g_{j,\alpha\beta})\prod_{m\in\mathcal{C}_i-\{j\}}[1-g_m] + \\
&\qquad 2\sum_{j,n\in\mathcal{C}_i}(g_{j,\alpha}g_{n,\beta}+g_{j,\beta}g_{n,\alpha}-g_{j,\alpha}g_{n,\alpha}-g_{j,\beta}g_{n,\beta})\prod_{m\in\mathcal{C}_i-\{j,n\}}[1-g_m]\Bigg)
\end{aligned}$$

where we have used the fact that partial derivatives commute for a convex function. For a convex function, g_j, we also have $g_{j,\alpha\alpha} \geq g_{j,\alpha\beta}$; it is trivial to show that when combined with the second condition this ensures that $f \geq 0$. Hence for any q_a, q_b and $|\delta| \leq \min(q_\alpha, q_\beta, c_\alpha - q_\alpha, c_\beta - q_\beta)$ we have $\mathbb{E}(PD|q_\alpha+\delta, q_\beta-\delta) \geq \mathbb{E}(PD|q_\alpha, q_\beta)$ for at least one sign of δ. The restriction on δ ensures that we do not spend a negative amount on a species or exceed its maximum expenditure.

For a fixed budget allocation to species α and β we have $q_\alpha + q_\beta = b$. By fixing the budget allocation, $\mathbb{E}(PD|q_\alpha, q_\beta)$ becomes a function of one variable

– either q_α or q_β – which we denote by $\mathbb{E}(PD|q_\alpha)$ and for which we have shown that

$$\mathbb{E}(PD|q_\alpha + \delta) \geq \mathbb{E}(PD|q_\alpha)$$

and/or

$$\mathbb{E}(PD|q_\alpha - \delta) \geq \mathbb{E}(PD|q_\alpha).$$

The remainder of this proof shows that if a non-extreme solution is optimal then $\mathbb{E}(PD|q_\alpha)$ is constant and therefore the extreme solutions have equal value.

Consider a problem where all optimal solutions are non-extreme and one of these is $\hat{\vec{q}}$. There will be two species for which:

$$0 < \hat{q}_\alpha < c_\alpha \text{ and } 0 < \hat{q}_\beta < c_\beta \qquad \text{with } \hat{q}_\alpha + \hat{q}_\beta = b.$$

Due to the budgetary constraints the domain of $\mathbb{E}(PD|q_\alpha)$ is

$$\max(0, b - c_\beta) \leq q_\alpha \leq \min(c_\alpha, b).$$

As extreme solutions are not optimal $\mathbb{E}(PD|q_\alpha)$ must be less than $\mathbb{E}(PD|\hat{q_\alpha})$ at the end points of this domain, consequently there will exist some limits around $\hat{q}_\alpha$ which we denote by l_- and l^- for which:

$$\mathbb{E}(PD|l_-) < \mathbb{E}(PD|\hat{q}_a),\ \mathbb{E}(PD|l^-) < \mathbb{E}(PD|\hat{q}_a) \text{ and}$$

$$\mathbb{E}(PD|q_\alpha) = \mathbb{E}(PD|\hat{q}_\alpha) \quad \text{for } l_- \leq \hat{q_\alpha} \leq l^-.$$

Consider the point $q_\alpha = (l_- + l^-)/2$, using our previous result we must have:

$$\mathbb{E}(PD|q_\alpha + \delta) \geq \mathbb{E}(PD|q_\alpha)$$

for at least one sign and all magnitudes of δ within the constraints of the domain, hence:

$$\mathbb{E}(PD|l_-) \geq \mathbb{E}(PD|q_\alpha) \text{ or } \mathbb{E}(PD|l^-) \geq \mathbb{E}(PD|q_\alpha).$$

A contradiction, thereby proving the existence of an optimal extreme solu-

tion.

7.2 Scenario 1: Constant costs and variable probabilities

Consider restricting the NAP such that all species cost the same amount to conserve ($c_j = c$) and only conserved taxa survive ($a_j = 0$, $b_j = 1$). In this situation the NAP becomes equivalent to the problem of finding PD maximising sets. Previously (chapter 6, Theorems 7 and 8) we have seen that a greedy algorithm can be used to solve this problem and consequently this restricted variant of the NAP. We now extend this result to allow non zero survival probabilities in the absence of conservation ($a_j \neq 0$). We retain the constraints that $b_j = 1$, c_j is constant and require the tree to be rooted.

Theorem 11. *Consider a NAP on a rooted tree where conservation costs are equal for all taxa ($c_j = c$) and conserved taxa are guaranteed to survive ($b_j = 1$). In this situation the algorithm in Theorem 7 can produce optimal solutions when applied to a rooted tree with suitably adjusted edge lengths, λ'_e. Denoting the set of children of edge e (the leaves/taxa separated from the root by e) by $\mathcal{C}_e$, the adjusted edge lengths are:*

$$\lambda'_e = \lambda_e \prod_{j \in \mathcal{C}_e} (1 - a_j). \tag{7.5}$$

Proof. Instead of maximizing $\mathbb{E}(PD|S)$ we can seek to maximize $\mathbb{E}(PD|S) - \mathbb{E}(PD|\emptyset)$, the increase in the expected PD that conservation of the taxa in S will provide. For our constrained NAP the increase in the probability that a particular edge is spanned when the set, S, of taxa is conserved is:

$$\begin{aligned} p(e|S) - p(e|\emptyset) &= \begin{cases} 1 - (1 - \prod_{j \in \mathcal{C}_e}(1 - a_j)), & \text{if } |\mathcal{C}_e \cap S| > 0; \\ 0, & \text{if } |\mathcal{C}_e \cap S| = 0; \end{cases} \\ &= \prod_{j \in \mathcal{C}_e} (1 - a_j) \times \begin{cases} 1, & |\mathcal{C}_e \cap S| > 0; \\ 0, & |\mathcal{C}_e \cap S| = 0. \end{cases} \end{aligned}$$

The expected increase in the PD is simply the sum over all edges with each edge weighted by the increased probability:

$$\begin{aligned}
\mathbb{E}(PD|S) - \mathbb{E}(PD|\emptyset) &= \sum_e \lambda_e (p(e|S) - p(e|\emptyset)) \\
&= \sum_e \lambda_e \prod_{j \in \mathcal{C}_e} (1 - a_j) \times \begin{cases} 1, & \text{if } |\mathcal{C}_e \cap S| > 0; \\ 0, & \text{if } |\mathcal{C}_e \cap S| = 0; \end{cases} \\
&= \sum_e \lambda'_e \times \begin{cases} 1, & \text{if } |\mathcal{C}_e \cap S| > 0; \\ 0, & \text{if } |\mathcal{C}_e \cap S| = 0. \end{cases}
\end{aligned}$$

This final expression for $\mathbb{E}(PD|S) - \mathbb{E}(PD|\emptyset)$ is equal to the objective, $\mathbb{E}(PD|S)$, for a Scenario 1 problem with branch lengths λ'_e as required.

This result can be extended further to permit all survival probabilities to be varied from unity / zero, however this requires a strict relationship between the a_j and the b_j values:

Theorem 12. *For the Noah's Ark Problem with equal conservation costs optimal solutions can be produced by a greedy algorithm if the following condition is met by the survival probabilities:*

$$\frac{1 - b_j}{1 - a_j} = \kappa, \tag{7.6}$$

for some constant κ (with $0 \leq \kappa \leq 1$).

The algorithm begins with an empty set S and sequentially adds the taxon, j, which maximizes $\mathbb{E}(PD|S \cup j)$ until S is at the maximum size permitted by the budgetary constraint.

Note that, if conservation is completely efficient ($b_j = 1$), the survival probabilities in the absence of conservation (a_j) are free to vary, otherwise this condition states that the extinction probability must be reduced by the same proportion for each taxon when it is conserved ($1 - b_j = \kappa(1 - a_j)$).

Proof. The proof proceeds in a similar fashion to Steel (2005) by establishing the strong exchange property discussed in section 6.1.2: namely that for any two subsets, Y and Z, of X with $|Y| < |Z|$ there exists some taxon $z \in Z$

such that:

$$\mathbb{E}(PD|Z - \{z\}) - \mathbb{E}(PD|Z) + \mathbb{E}(PD|Y \cup \{z\}) - \mathbb{E}(PD|Y) \geq 0. \quad (7.7)$$

This means that for any two subsets of X, the larger subset contains some taxon (z) that would contribute more to the expected PD value of the smaller subset than it adds to that of the larger one.

Denote the set of edges on the path from z to the root by R, and notice that each of the expected PD terms in (7.7) can be split into a sum over the edges in R, and a sum over the edges not in R. The significance of this observation is that the probability that edges not in R are spanned remains unchanged as z is removed from Z or added to Y. Denoting the left hand side of Equation 7.7 by ΔPD we have:

$$\Delta PD = \sum_{i \in R} \lambda_i \Delta p(i) + \sum_{j \notin R} \lambda_j \Delta p(j),$$

where

$$\Delta p(i) := p(i|Z - \{z\}) - p(i|Z) + p(Y \cup \{z\}) - p(i|Y),$$

then for $j \notin R$ we have $\Delta p(j) = 0$ since the probability of an edge not in R being spanned is independent of the presence of taxon z, hence:

$$\Delta PD = \sum_{i \in R} \lambda_i \Delta p(i).$$

A sufficient condition for satisfying the strong exchange property (Equation 7.7) is therefore that $\Delta p(i) \geq 0$ for each edge i on the path from taxon z to the root. The following results follow from the definition of $p(i|Z)$:

$$\begin{aligned} p(i|Z - \{z\}) - p(i|Z) &= (a_z - b_z) \prod_{m \in \mathcal{C}_i - Z} (1 - a_m) \prod_{l \in \mathcal{C}_i \cap Z - \{z\}} (1 - b_l) \\ p(i|Y \cup \{z\}) - p(i|Y) &= (b_z - a_z) \prod_{m \in \mathcal{C}_i - Y - \{z\}} (1 - a_m) \prod_{l \in \mathcal{C}_i \cap Y} (1 - b_l). \end{aligned}$$

Combining these gives an identity for $\Delta p(i)$ which can be further simplified

using Equation 7.6 :

$$\begin{aligned}\Delta p(i) &= \Big(\prod_{m \in \mathcal{C}_i - Z} (1 - a_m) \prod_{l \in \mathcal{C}_i \cap Z - \{z\}} (1 - b_l) \\ &\quad - \prod_{m \in \mathcal{C}_i - Y - \{z\}} (1 - a_m) \prod_{l \in \mathcal{C}_i \cap Y} (1 - b_l) \Big) (a_z - b_z) \\ &= \left(\kappa^{|\mathcal{C}_i \cap Z| - 1} - \kappa^{|\mathcal{C}_i \cap Y|}\right) (a_z - b_z) \prod_{m \in \mathcal{C}_i} (1 - a_m)\end{aligned}$$

Noting that $a_z - b_z$ is negative, a sufficient condition for insuring that $\Delta p(i) \geq 0$ is $\kappa^{|\mathcal{C}_i \cap Z| - 1} - \kappa^{|\mathcal{C}_i \cap Y|} \leq 0$, which (since $0 \leq \kappa \leq 1$) is equivalent to

$$|\mathcal{C}_i \cap Y| \leq |\mathcal{C}_i \cap Z| - 1. \tag{7.8}$$

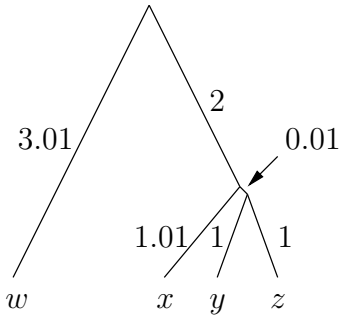
This condition simply states that the number of elements in Y that span edge i, is strictly less than the number of elements in Z that span that edge.

Next we show that for any two sets Y and Z with $|Y| < |Z|$ it is possible to find a taxon z for which this last property holds for each edge i on the path from z to the root.

Starting at the root, one of the edges adjacent to the root must satisfy Equation 7.8 since $|Y| < |Z|$ (if Equation 7.8 were not satisfied this would imply $|Y| \geq |Z|$), call this edge m. Similarly one of the edges below m must satisfy Equation 7.8 since we have $|\mathcal{C}_m \cap Y| \leq |\mathcal{C}_m \cap Z| - 1$, pick this edge, call it m and continue this procedure until one arrives at an exterior edge. The condition $\Delta p(i) \geq 0$ is therefore met on every edge, from the taxon adjacent to this exterior edge through to the root, consequently the strong exchange property (Equation 7.7) holds.

Let Y be an optimal solution if m taxa are to be conserved and Z an optimal solution if $m + 1$ taxa are to be conserved. Applying the strong exchange property (Equation 7.7) to Y and Z shows the existence of a taxon z such that $Y \cup \{z\}$ is an optimal solution for $m + 1$ taxa and $Z - \{z\}$ is an optimal solution for m taxa.

Theorem 12 follows easily by standard arguments from 'greedoid' theory (Korte et al., 1991). Specifically the above observation shows that any solution for $m + 1$ taxa must be obtained from a solution for m taxa by adding



Subset (S)	$\mathbb{E}(PD\|S) - \mathbb{E}(PD\|\emptyset)$
$\{w\}$	0.602
$\{x\}$	1.068
$\{y\}$,$\{z\}$	1.055
$\{w, y\}$,$\{w, z\}$	1.657
$\{w, x\}$	1.670
$\{y, z\}$	1.745
$\{x, y\}$, $\{x, z\}$	1.672

Figure 7.1: A NAP that does not satisfy condition 7.6 and violates the substructure property. The optimal subset of size 1 is $\{x\}$ and the optimal subset of size 2 is $\{y, z\}$. Parameter values are $a_w = 0.6$, $a_x = 0.5$, $a_y = a_z = 0.25$, $b_w = 0.8$, $b_x = 1$ and $b_y = b_z = 0.85$.

a single taxon which maximizes the increase in $\mathbb{E}(PD|Y)$.

7.2.1 The Necessity of Equation 7.6

Any problem for which the greedy algorithm is optimal must satisfy the 'substructure' property (Cormen et al., 2002). This property states that an optimal solution, Y, of a given size must be contained within an optimal solution of each larger size. The condition imposed in the previous section (Equation 7.6) ensures that the substructure property holds for the optimization problem.

Here we provide a simple example to show that this substructure property (and thereby the greedy algorithm) can fail when the condition imposed by Eqn. 7.6 in Theorem 12 is violated.

In Figure 7.1 the optimal subset of size 1 is $\{x\}$. The additional contribution to $E(PD)$ made by the pendant edge of x when it is conserved is smaller than that from the pendant edges of y or z (were they to be conserved). The optimality of x is entirely due to its conservation ensuring that the interior edge of length 2 is spanned.

When two taxa are conserved the probability increase that x provides for the interior edge of length 2 is reduced such that the smaller increase in this probability that y and z provide. Coupled with the greater contribution from their pendant edges makes x a less valuable taxon to conserve. The

optimal subset of size 2 is therefore $\{y, z\}$ (see Figure 7.1), the substructure property is violated (which was possible as the condition in the previous section (Equation 7.6) was not satisfied) and the greedy algorithm cannot produce the optimal solution.

7.2.2 Non-linear expenditure-survival relationship

Recall that the expenditure-survival relationship $g_j(q_j)$ gives the probability that a taxon, j, will remain extant given that q_j is spent on its conservation. Scenario 1 and Scenario 2 (in the following section) assume a linear relationship for $g_j(q_j)$ (Equation 7.3). This linear relationship ensures that solutions are extreme – all taxa with one possible exception are fully conserved or not at all – which in turn simplifies the NAP problem from one of deciding the amount to spend on the conservation of each taxon to one of selecting the optimal set of taxa to conserve.

Simianer et al. (2003) questioned the validity of the linear relationship and applied the NAP using various alternatives to Equation 7.3. Further examples of different relationships can be found in Johst et al. (2002) and Lamberson et al. (1992).

For convenience, problems with $g_j(q_j)$ not of the type given in Equation 7.3 will be referred to as 'Generalised Noah's Ark Problems' (g-NAPs). The relationships $g_j(q_j)$ within a g-NAP are generally not parametrized by a_j, b_j and c_j and cannot be assumed to have extreme solutions. However, as we will show, there is one family of g-NAPs which can be solved using a greedy algorithm, namely g-NAPs where $g_j(q_j)$ has the form:

$$g_j(q_j) = 1 - k^{q_j}(1 - a_j) \text{ with } 0 \leq k \leq 1,$$

can be transformed to a NAP (with $g_j(q_j)$ as in Equation 7.3) using the method detailed in section 7.4; the resulting NAP is of the type described in Scenario 1. Consequently such problems can be solved using a greedy algorithm.

This formulation of $g_j(q_j)$ corresponds to the situation where each budgetary unit allocated to conserving a taxon produces progressively smaller increases in that taxon's survival probability as dictated by the above equa-

tion. Note that survival of a taxa cannot be guaranteed regardless of the funding allocated to its conservation (unless of course $a_j = 1$).

Other g-NAPs that satisfy certain conditions (discussed in Appendix 7.4) can be transformed to NAPs. The resulting NAPs will generally not fall into Scenario 1 and may therefore violate the substructure property, hence they may not be solvable using a greedy algorithm.

7.3 Scenario 2: Variable Conservation Costs and a Ultrametric Tree

In this section a variation of the NAP is considered that allows variable conservations costs and for which the greedy algorithm can produce an optimal solution (W).

Denote the expected contribution a particular taxon, j, makes to the expected PD of a set of taxa, W, by $\sigma_W(j)$. That is, if j is in W, $\sigma_W(j)$ is the PD that W would lose if j were removed, if j is not in W it is the PD that W can gain from the addition of j:

$$\sigma_W(j) := \mathbb{E}(PD|W \cup \{j\}) - \mathbb{E}(PD|W - \{j\}).$$

The cost-benefit of adding a taxon to a subset is given by $r_W(j) = \sigma_W(j)/c_j$, this is the contribution j makes to the PD per unit of cost. The overall cost benefit of a particular subset of taxa W is $R_W = \mathbb{E}(PD|W)/\sum_{j\in W} c_j$, and optimal solutions to the NAP will maximize R_W subject to the total cost equaling the conservation budget (B).

Theorem 13. *A greedy algorithm produces optimal solutions for any Noah's Ark Problem with variable conservation costs provided the tree is ultrametric and conservation increases the survival probability of each taxon from certain extinction ($a_j = 0$) to certain survival ($b_j = 1$).*

The greedy algorithm begins with $W = \emptyset$ and continues to add the taxon with the highest value of $r_W(i)$ to W until the cost of conserving the taxa in W exceeds the budget. The last taxon added should be partially conserved to bring the total cost to the budget.

This theorem is a variation of that stated, without reference or proof, in

Weitzman (1992) (page 374) and Weitzman (1995) (page 31). The difference between the proposed algorithms is that the greedy algorithm presented here builds up a set of taxa to conserve by adding one taxon at a time whereas that proposed by Weitzman begins with the full set of taxa and removes one taxon at a time. The requirement in Weitzman (1992) that the dissimilarity measure be ultrametric and the requirement in Weitzman (1995) of a bead model of evolutionary branching are both equivalent to requiring the tree to be ultrametric. Weitzman's theorem claims that the greedy algorithm will produce optimal results for an ultrametric tree and it allows for intrinsic values of the conserved taxa (as discussed previously). However it is the modified tree where the intrinsic values of the taxa have been added to the pendant edges that must be ultrametric.

Proof. Theorem 13 cannot be proven in the same manner as Theorem 12 because the strong exchange property (Equation 7.7) does not hold (it is a straightforward matter to construct a counterexample). Instead, for this scenario we establish two claims: (i) all subsets not produced by the greedy algorithm are sub-optimal and (ii) that all subsets produced by the greedy algorithm produce subsets of the same value.

Claim (i). Suppose that W is an optimal subset that can not be produced by the greedy algorithm. Consider constructing W by beginning with an empty set and adding the elements in W one at a time such that a greedy choice is made whenever possible. Since W cannot be produced by a greedy algorithm there will be some point in this sequence where a taxon, h is added instead of a greedy choice, denote the subset to which h is added by Y ($Y \subset W$) and a taxon that the greedy algorithm would have added by g ($g \in W, g \notin Y$).

Consider the taxon in W that is the closest to g, without loss of generality the situation is as depicted in Figure 7.2. Denote this taxon by j (this taxon may not be unique, in this case the choice of j is arbitrary). It is necessary to consider two cases: $j \in W - Y$ and $j \notin W - Y$.

If $j \in W - Y$, g was a greedy choice at a time where j could have been

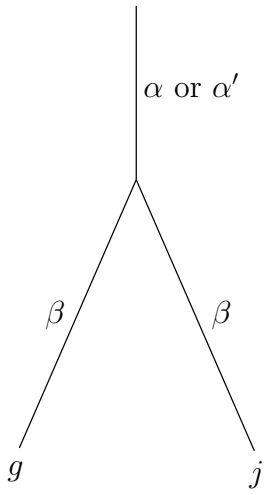


Figure 7.2: The general situation when two taxa, g and j, that share a common edge not in $\mathcal{T}|Y$ are added to $\mathcal{T}|W$. The tree has been assumed to be ultrametric. The root of the depicted tree corresponds to an interior node of $\mathcal{T}|Y$ or $\mathcal{T}|W$ when the length of the root edge is α or α' respectively.

added to the subset Y and since the greedy choice was not made we have:

$$\begin{array}{rcll} r_Y(g) & > & r_Y(j), & \text{that is,} \\ \dfrac{\alpha+\gamma}{c_g} & > & \dfrac{\alpha+\gamma}{c_j} & \text{(using the branch lengths in Figure 7.2)} \\ c_g & < & c_j. & \end{array} \tag{7.9}$$

The cost benefits of g and j relative to the final subset (W) are:

$$\begin{aligned} r_W(g) &= (\alpha'+\gamma)/c_g, \\ r_W(j) &= (\alpha'+\gamma)/c_j. \end{aligned}$$

From Equation 7.9 we have $c_g < c_j$, hence $r_W(g) > r_W(j)$. The cost benefit of g exceeds that of j; diverting some funding from taxon j to g will increase the overall cost benefit, hence W is not an optimal subset.

If $j \notin W - Y$ there is no taxon in $W - Y$ that can reduce the cost benefit of g, hence the cost benefit of g still exceeds that of h and diverting funding from h to g will again increase the overall cost benefit, hence W is not an optimal subset. Hence all optimal solutions must be produced by a greedy

algorithm.

Claim (ii). Since an optimal solution exists (but may not be unique) at least one solution produced by the greedy algorithm must be optimal. To show that all solutions are in fact optimal it suffices to examine what happens when the greedy algorithm has to select from several greedy choices to add to a subset Y. Consider the case where there are two taxa, j and k with equal cost benefit. This can occur in two ways as depicted in Figure 7.3.

Case 1. The taxa with equal cost benefit attach to different internal nodes of $\mathcal{T}|Y$. In this case addition of either taxa does not effect the cost benefit of the other taxon; regardless of which taxon is conserved first the other will be conserved next at the same cost benefit.

Case 2. The taxa with equal cost benefit attach to the same internal node of $\mathcal{T}|Y$. This situation is more complex, addition of the first taxon reduces the cost benefit of the second taxon consequently other taxa may have a higher cost benefit and be conserved before the second taxon.

As j and k have the same cost benefit the fact that the tree is ultrametric dictates that j and k have the same cost. It is therefore apparent that both the remaining budget and the cost benefit of the unconserved taxa are independent of which of j and k is conserved first. Only the cost benefits of those taxa that are incident with the pendant edge of j or k in $\mathcal{T}|Y \cup \{j, k\}$ (for example taxon m in Figure 7.3) are dependent on which of j and k is first conserved. However from the same argument used to produce Equation 7.9 all of these taxa will have a higher cost than j and k, subsequently they will not be conserved until both j and k have been conserved (at which point it becomes irrelevant which of these taxa was conserved first).

The extension to more than two taxa with the same cost benefit, possibly with combinations of these two scenarios is straightforward.

7.3.1 Beyond Ultrametric Trees

When applied to a tree that is not ultrametric the greedy algorithm is no longer guaranteed to provide the optimal solution. In particular when new taxa are added by the greedy algorithm it is possible for taxa that have been added previously to have their cost benefit reduced below that of some

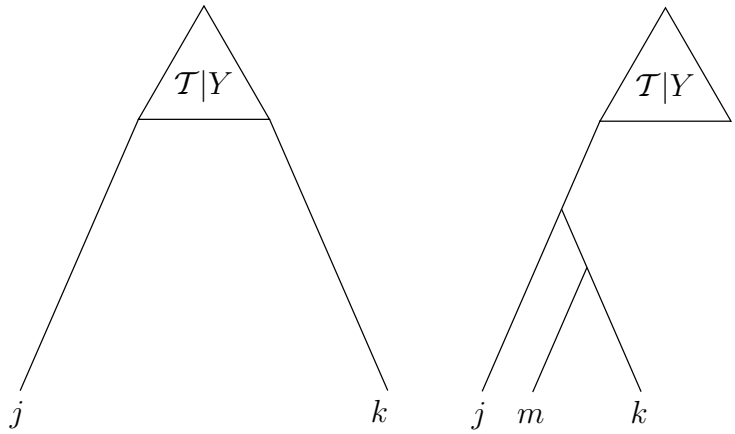


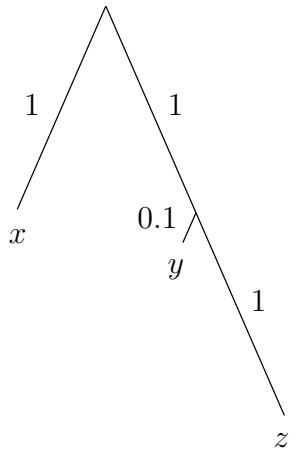
Figure 7.3: The two ways in which taxa with the same cost benefit may attach to an existing tree, $\mathcal{T}|Y$. Note that in both cases there may be any number of other taxa not in Y that attach to the edges depicted (such as the taxon m).

taxon not selected thus far - this problem may not exhibit the substructure property. This is illustrated in Figure 7.4. The optimal subset of one size is $\{y\}$ with a cost benefit of 1.1, whereas the optimal subset of size two is $\{x, z\}$ with a cost benefit of 0.71.

Note that this problem is equivalent to a problem where the pendant edges of y and z have zero length and x, y and z have intrinsic values of 0, 0.1 and 1 respectively. The resulting tree is ultrametric and thus by the theorem proposed in Weitzman (1992) (page 374) and Weitzman (1995) (page 31) should be solveable by their greedy algorithm. However since the optimal solutions do not satisfy the substructure property they cannot be produced by any greedy algorithm. If intrinsic values are being considered the tree formed when these values are added to the pendant edges must be ultrametric for a greedy algorithm to produce optimal solutions.

7.4 Scenario 3: The generalised Noah's Ark Problem

We describe a technique by which Generalised Noah's Ark Problems (g-NAPs) that satisfy certain conditions are transformed to equivalent NAPs. This transformation is used to show that there is one form of $g_j(q_j)$ that



Subset (S)	Overall Cost Benefit (W_S)
$\{x\}$	0.5
$\{z\}$	0.89
$\{y\}$	1.1
$\{y, z\}$	0.65
$\{x, y\}$	0.70
$\{x, z\}$	0.71

Figure 7.4: A tree that is not ultrametric can lead to a violation of the substructure property. The optimal subset of size 1 is $\{y\}$ and the optimal subset of size 2 is $\{x, z\}$. Parameter values are $c_x = 2$, $c_y = 1$, $c_z = 2\frac{1}{4}$ and edge lengths as indicated.

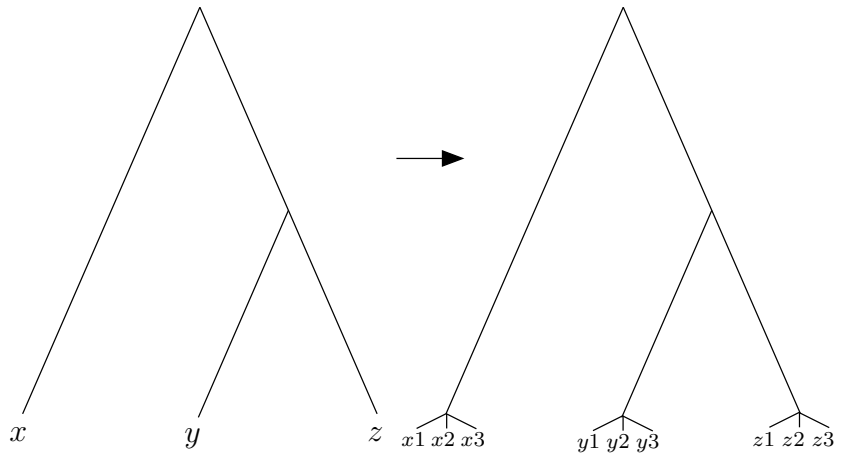


Figure 7.5: The transformation applied to the simple tree on the left. In this example $m = 3$ such that each taxon in the original g-NAP is replaced by three derived taxa with pendant edges of zero length. Each taxon, o, derived from taxon j, will have $c_{jo} = \delta$ and, a_{jo} and b_{jo} that satisfy Equation 7.10 for all l.

transforms to the type of problem considered in Scenario 1 and can therefore be solved using a greedy algorithm.

We may assume that there is some smallest unit by which the q_j can be increased or decreased (the absolute limit is the smallest unit of currency), and we denote this by δ. Recalling that the conservation budget is B, there are $m = B/\delta$ units of budget to allocate. In the transformed problem each taxon, j, from the original g-NAP is replaced by m derived taxa (see Figure 7.5). The m derived taxa are all located in the same position in the tree as the original taxon j was, this is possible as these taxa have pendant edges of zero length and the original taxon j is a leaf node.

Each of the derived taxa represents a budget unit being allocated to the original taxon j. Consequently there is an ordering of these taxa, the first of the m taxa derived from j represents a single budget unit being allocated to j and so on. Given a solution to the transformed NAP the corresponding solution to the g-NAP is found by noting how many derived taxa are conserved for each original taxon, j – this indicates the number of budgetary units to allocate to j.

The cost of each derived taxon is simply the cost of a single budgetary unit (δ). Next it is necessary to place some restrictions on the parameters, a_{jl} and b_{jl} of the derived taxa. Consider a taxon, j, in the original g-NAP. When the first l taxa derived from j are conserved the probability that at least one of the taxa derived from j remains extant is:

$$z_{jl} = 1 - \prod_{o \leq l}(1 - b_{jo}) \prod_{r > l}(1 - a_{jr}).$$

For the derived NAP to be equivalent to the original g-NAP z_{jl} should equal the probability that j remains extant if $l\delta$ is spent on conserving it: $q_j(l\delta)$. For each original taxon, j, this gives $m + 1$ equations for the $2m$ parameters b_{jl} and a_{jl}:

$$z_{jl} = q_j(l\delta). \tag{7.10}$$

Lemma 1. *The above transformation results in a NAP that is equivalent to the original g-NAP provided that for all j and for all l:*

$$\frac{b_{j(l+1)} - a_{j(l+1)}}{1 - a_{j(l+1)}} \leq \frac{b_{jl} - a_{jl}}{1 - a_{jl}} \tag{7.11}$$

Proof. From the derivation of the condition on a_{jl} and b_{jl} it is apparent that conserving the first l taxa derived from the original taxon j is equivalent to spending δl on conserving taxon j. However this assumes that the derived taxa are added in the appropriate order, the remainder of this proof shows that this is guaranteed if Equation 7.11 is satisfied.

Consider only those taxa derived from a single taxon, j, of which the first l taxa in the sequence have been conserved. The increase in z_{jl} that the addition of one of the remaining taxa, o, will provide is:

$$\Delta z_{jl}(o) = \frac{b_{jo} - a_{jo}}{1 - a_{jo}} \prod_{r \leq l} (1 - b_{jr}) \prod_{s > l} (1 - a_{js}).$$

The taxon that provides the greatest increase in z_{jl} will be the taxon picked next by the greedy algorithm. Equation 7.11 guarantees that $\Delta z_{jl}(o)$ will be greatest for $o = l + 1$, hence the correct taxon may be added next. There may be other taxa with an equal value of $\Delta z_{jl}(o)$ however it is only necessary for the correct sequence of taxon additions to be a possible greedy solution. As previously noted all solutions produced by the greedy algorithm will be optimal, hence it suffices for one of the solutions produced by the transformed NAP to be realistic.

Theorem 14. *Problems for which $g_j(q_j)$ has the form*

$$g_j(q_j) = 1 - k^{q_j}(1 - a'_j) \text{ with } 0 \leq k \leq 1, \tag{7.12}$$

can be transformed to a NAP of the type described in Scenario 1. Consequently such problems can be solved using a greedy algorithm.

Proof. To satisfy the restrictions imposed on Scenario 1 the costs of each transformed taxon must be equal and Equation 7.6 must be satisfied. The former restriction is trivial as each taxon costs δ to conserve, the remainder of the proof shows that a transformation satisfying the latter condition exists.

The condition imposed on the transformation ($z_{jl} = g_j(l\delta)$) for this particular $g_j(q_j)$ is:

$$1 - \prod_{o \leq l}(1 - b_{jo}) \prod_{r > l}(1 - a_{jr}) = 1 - k^{l\delta}(1 - a_j). \tag{7.13}$$

Applying the necessary condition for the transformed NAP to be a Scenario 1 type problem (Equation 7.6) this becomes:

$$1 - \kappa^l \prod_{r}(1 - a_{jr}) = 1 - k^{l\delta}(1 - a_j). \tag{7.14}$$

This has a simple solution, $\kappa = k^\delta$ and $a_{jr} = 1 - (1 - a_j)^{1/m}$ for all j, r. This solution also trivially satisfies Equation 7.11 since all taxa derived from an original taxon are identical (and hence the transformed NAP is equivalent to the original g-NAP).

7.5 Concluding comments

Simple greedy algorithms were outlined for two special cases of the Noah's Ark Problem (NAP) and an extension of the NAP – the g-NAP. These special cases are more realistic than that considered by Steel (2005) for which it is known that a greedy algorithm exists. Using these algorithms optimal solutions for practical problems that fall within these scenarios can be computed efficiently.

Simianer et al. (2003), (page 384) has suggested (without proof) that a greedy algorithm will produce optimal solutions for a family of problems equivalent to the Generalised Noah's Ark Problem (g-NAP) described here, provided the tree satisfies a molecular clock. This family of problems includes the NAP proposed by Weitzman (1998) for which we have illustrated several cases where a greedy algorithm cannot produce optimal solutions (Figures 7.1 and 7.4). Hence we have shown that greedy algorithms are not, in general, guaranteed to produce optimal solutions for NAPs or g-NAPs. Caution is advised when applying a greedy algorithm to a problem not of the types described in Scenarios 1 and 2 - the solutions produced may not be optimal.

Reist-Marti et al. (2006) describe a two step algorithm for solving g-NAPs, they note that this algorithm is not guaranteed to produce the optimal solution. Algorithms such as this may prove useful, particularly for

more complicated variations of the NAP. It would also be of interest to determine how close the solutions produced by such algorithms are to the global optimal.

The Noah's Ark Problem provides a satisfying framework for biodiversity resource allocation problems. It is however, still a simplification of reality and some extensions to it have been suggested.

The NAP as presented here does not consider the possibility of partially conserving taxa and therefore being able to spread resources more thinly across a greater number of taxa. Weitzman (1998) assumed that the survival probability of a taxon increases linearly with the conservation funding allocated to that taxon. Under this assumption optimal solutions to the NAP are extreme and allocate the maximum possible amount to a few taxa instead of partially conserving a greater number. An extension of the NAP to more realistic relationships between survival probability and expenditure was considered in Simianer et al. (2003) with an application to conservation of breed diversity in African cattle. A greedy algorithm was presented in that paper that the authors suggested would provide optimal solutions to all problems of this type. However it was shown in Hartmann and Steel (2006) that this cannot be the case. This was extended further in Reist-Marti et al. (2006) to allow for discontinuous relationships produced by multiple possible conservation schemes, necessitating a two step optimisation procedure (which they state is not guaranteed to produce the global optimum).

Another implicit assumption in the NAP is that the survival probabilities are independent. That is, conserving one taxon does not raise or lower the survival probabilities of any others and this may be unrealistic. For example, conserving the prey of one taxon may raise the survival probability of that taxon as well. This effect was considered in van der Heide et al. (2005) where it was shown that failure to consider interdependent survival probabilities may result in an incorrect suggestion as to which species should be protected. The authors in this study stress the importance of their findings as "more significant losses of biodiversity are exactly those in which ecological impacts are severe, that is, where the loss of one species affects the survival of others".

In summary, whilst the NAP provides a good starting point, there are other important factors that influence which taxa should be conserved. In-

clusion of some of these may prove more difficult than others and adding these factors will further complicate the problem of finding optimal solutions. For example, consider the following problem which is relevant to biodiversity conservation. We have a collection C of locations, where each location $l \in C$ contains some subset $S(l)$ of taxa from a set X of taxa; also we have a phylogenetic X–tree $\mathcal{T}$ with branch lengths. We wish to select k locations so as to maximize the PD of the set of taxa that occur in at least one selected location. If no taxon occurs in more than one location this problem is easily solved, by transforming it to the standard PD optimization problem and applying the greedy algorithm. In general, however, the problem is NP–hard. The proof consists of showing that one can transform the NP–complete problem 'Minimum cover' (Garey and Johnson, 1979) to this problem, by selecting branch lengths for $\mathcal{T}$ that are 1 on all the pendant edges, and 0 on all the interior edges. For various approaches to solving this and related problems see Rodrigues et al. (2005), Camm et al. (2006) and Wilson et al. (2006).

Chapter VIII

The Noah's Ark Problem with uncertain data

One of the criticisms of the NAP is that the input parameters (the survival probabilities, conservation costs and phylogenetic information) are difficult to determine and often only simple estimates will be available. In this chapter we consider uncertainty in the survival probabilities of unconserved species. This is arguably the parameter with the greatest uncertainty and fortunately one that is easy to handle mathematically. To consider this uncertainty we develop an extension of the NAP – the uncertain NAP (uNAP). The uNAP explicitly takes uncertainty about the survival probabilities into account to produce solutions that are robust across the range of possible probabilities.

To investigate the uNAP we consider two datasets, i) trees sampled from the Yule model with arbitrary survival probabilities and ii) a newly constructed tree of Madagascan lemurs with extinction risk categories as obtained from IUCN (2007). Solutions obtained with the NAP and the uNAP provided significant improvements over random species choice or simple PD maximising sets, for both datasets. This shows that simple estimates of survival probabilities add much information, even if these point estimates are poor. This is particularly relevant as many current approaches have sought to simply maximise PD.

Surprisingly, but reassuringly, the uNAP provided solutions that were only marginally better than solutions obtained using the NAP. This indicates that the NAP is robust to uncertainty in the survival probabilities. Furthermore this suggests that in some situations the gain provided by the uNAP over the NAP may be insufficient to justify the additional complexity of the uNAP.

8.1 Uncertain Survival Probabilities

In previous chapters we assumed that the input parameters for the NAP are known with certainty, for real situations this is clearly a questionable assumption. In this chapter we consider the effect of this assumption for the survival probabilities of species in the absence of conservation (a_j). There are two reasons why we consider uncertainty in these survival probabilities. Firstly from a mathematical perspective these are the simplest parameters for which to include uncertainty. Secondly, depending on the conservation scenario these are arguably the parameters with the highest uncertainty.

Uncertainty in the survival probabilities for unconserved species comes from two main sources. The first of these is due to the difficulty in determining the relative survival probabilities of all the species – how much more likely is it that one species will survive than another? The second source of uncertainty as discussed in Hartmann and Steel (2007) is due to the fact that the survival probabilities implicitly determine a management time scale. The survival probabilities should be interpreted as the probability that a species will survive to a particular time in the future. If this time is far into the future the overall magnitude of the survival probabilities will be low (as all species eventually become extinct) whereas if this time is closer to the present the overall magnitude will be higher. In the following section we show that the optimal solution to the NAP is dependent on the overall magnitude of the survival probabilities.

A discussion of appropriate time scales / survival probability magnitudes is beyond the scope of this chapter. For example, if management is flexible such that resources can be reallocated on a short time scale, the survival probabilities should be calculated for a time closer to the present. This is because management will be able to respond (and reallocate resources) to any extinctions or drastic changes in the NAP parameters that occur.

8.2 Conservation Timescale

The survival probabilities in the NAP (a_j) contain an implicit time scale as they represent the probability that a taxon will survive to some future time,

t. In the absence of conservation the expected number of taxa surviving to t is $\sum_j a_j$. If the time t is in the distant future (a long time scale) the survival probability of unprotected taxa will be close to zero due to background extinction. For shorter time scales (t closer to the present) the survival probabilities will be closer to one. This choice of time scale affects solutions to the NAP as management strategies corresponding to longer time scales will place greater emphasis on internal edges.

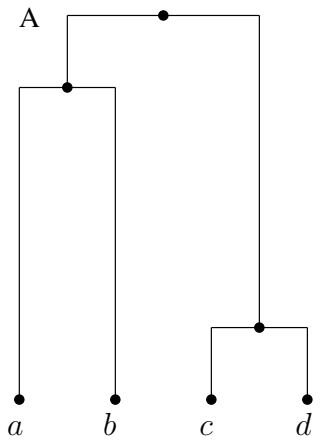
To illustrate the importance of selecting an appropriate time scale consider the tree in Figure 8.1 where each taxon is equally likely to remain extant at any future time. Panel A corresponds to the situation where all taxa that are not conserved become extinct (a long time scale). If two taxa can be conserved the optimal choice consists of one taxon from each branch of the tree. This optimal choice is found either by application of the greedy algorithm (Theorem 7) or by an exhaustive search.

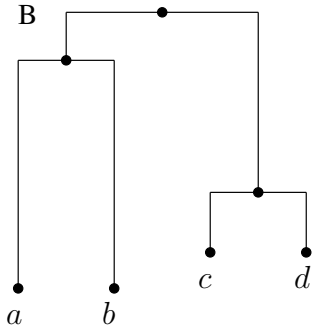
Consider increasing the survival probability of unconserved taxa (a_j) so that all taxa have a $\frac{1}{4}$ chance of surviving; this represents a move to a shorter management time scale. To find the optimal solutions for this problem the transformation outlined in Theorem 11 is applied to the original tree (Panel A in Figure 8.1) yielding the tree in Panel B. As expected from Equation 7.5 the interior edges have had a greater reduction in length than the pendant edges; application of the greedy algorithm can now be used to obtain the optimal solutions. The pendant edge lengths of taxa a and b are now equal to the distance between the root and taxa c or d. Consequently conserving both taxa a and b is now also an equally good solution.

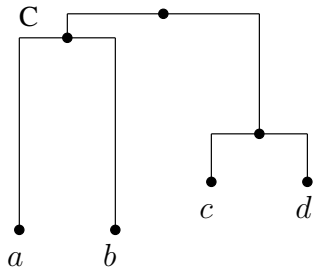
If the survival probabilities (a_j) are further increased (to, say, $\frac{1}{2}$), the interior edges of the transformed tree decrease in length to such an extent that the optimal set of taxa to conserve becomes $\{a, b\}$ (see Panel C).

We have illustrated that the optimal set of taxa to conserve is dependent on the management time scale. As the management time scale shifts from long term to short term less emphasis is placed on interior edges as these are more likely to remain extant anyway.

A discussion of the merits of conservation time scales is beyond the scope of this work (see Bunnell and Huggard (1999) and Lewis et al. (1996) for more details). However the optimal timescale will be highly dependent on







Conserved Taxa (S)	Optimal? A	B	C
$\{a, b\}$	✗	✓	✓
$\{c, d\}$	✗	✗	✗
$\{a, c\}$,$\{a, d\}$, $\{b, c\}$,$\{b, d\}$	✓	✓	✗

Figure 8.1: Panel A depicts a tree where unconserved species become extinct with certainty ($a_j = 0$). Panels B and C depict the trees transformed according to Theorem 11 as the survival probability is increased to 0.25 and 0.375 respectively. Optimal subsets of size two can be found by applying the greedy algorithm to these trees (Theorem 11). The optimality of each subset for each panel is indicated in the table.

the application. Of particular importance will be the time scale on which conservation focus can be shifted from one taxon to another. If this can occur rapidly, planning for the short term would be optimal and the conservation strategy should be reevaluated as taxa become extinct. For many taxa, conservation programs are long term investments. In these cases, a longer time scale should be investigated when the taxa to be conserved are initially selected.

8.3 Incorporating uncertainty

In this section we present an extension to the NAP which allows uncertainty about the survival probabilities to be incorporated in a rigorous manner. We will refer to this extension as the Uncertain-NAP (uNAP). In the uNAP a joint probability density for the survival probabilities is specified. For the species that are not conserved we denote the vector of survival probabilities by $\vec{a} = [a_1, a_2, \ldots, a_n^T]$ and the probability density by $\phi(\vec{a})$. Similarly the survival probabilities for conserved species have a density $\vec{b}$. These densities can be arbitrarily complicated and may include interactions between species. However the survival probability of a species cannot be made dependent on which other species are conserved. Taking this dependence into account would make the optimisation problem far more complicated. The probability density can also be arbitrarily simple such as uniform probability density over some interval or over all possible values (zero to one); such simple densities may be of more practical interest.

Previously we denoted the objective function for the NAP by $\mathbb{E}(PD|S)$. Equivalently we can maximise $\mathbb{E}(PD|S) - \mathbb{E}(PD|\emptyset)$ since the latter term is independent of S. We now make the dependence on the survival probabilities explicit such that the objective function is $\mathbb{E}(PD|S, \vec{a}, \vec{b}) - \mathbb{E}(PD|\emptyset, \vec{a}, \vec{b})$. Since there is uncertainty about these survival probabilities we need to take the expectation over $\vec{a}$ and $\vec{b}$, therefore the objective function for the uNAP is:

$$\mathbb{E}_{\vec{a},\vec{b}}(\mathbb{E}(PD|S, \vec{a}, \vec{b}) - \mathbb{E}(PD|\emptyset, \vec{a}, \vec{b})). \tag{8.1}$$

Using this objective function we can succinctly define the uNAP:

GIVEN AN EDGE-WEIGHTED PHYLOGENETIC TREE, CONSERVATION COSTS c_j FOR EACH TAXON AND JOINT SURVIVAL PROBABILITY DENSITIES $\vec{a}$ AND $\vec{b}$, MAXIMISE EQUATION 8.1 SUBJECT TO THE CONSTRAINT: $\sum_{j \in S} c_j \leq B$.

Finding solutions to the uNAP will be at least as complex as obtaining solutions to the NAP. Fortunately a greedy algorithm similar to that in Theorem 11 can be applied to a particular instance of the uNAP where all species cost the same to conserve and conserved species survive with certainty ($\phi(\vec{b}) = \delta(\vec{1})$). This is obviously a unrealistic restriction, however it vastly simplifies the analysis of uncertainty in $\vec{a}$ as solutions can readily be obtained. In some cases where conservation dramatically increases survival probabilities, the assumption of conserved species surviving with certainty may not be that unrealistic (eg. for captive breeding).

Theorem 15. *Consider an instance of the uNAP where all species cost the same to conserve ($c_j = c$) and all conserved species survive with certainty. Optimal solutions for such a problem can be obtained using a transformation and the greedy algorithm outlined in Theorem 7. The transformation involves rescaling the edges of the tree using:*

$$\lambda' = \lambda_i \int_{\vec{a}} \phi(\vec{a}) \prod_{j \in C_i} (1 - a_j) d\vec{a}, \tag{8.2}$$

Proof. The objective function for this instance of the uNAP can be expressed as

$$\mathbb{E}_{\vec{a}}(\mathbb{E}(PD|S, \vec{a}) - \mathbb{E}(PD|\emptyset, \vec{a}))$$

since $\vec{b}$ is fixed at unity (guaranteed survival). Making the outer expectation explicit we get:

$$\mathbb{E}_{\vec{a}}(\mathbb{E}(PD|S, \vec{a}) - \mathbb{E}(PD|\emptyset, \vec{a})) = \int_{\vec{a}} \phi(\vec{a}) \left(\mathbb{E}(PD|S, \vec{a}) - \mathbb{E}(PD|\emptyset, \vec{a})\right) d\vec{a}. \tag{8.3}$$

We denote the probability that an edge will 'survive' if a set S is conserved, by $p_S(i)$. If S contains a leaf below edge i then $p_S(i) = 1$ since all conserved species survive. If S does not contain a leaf below edge i, the probability that edge i will survive is one minus the probability that all leaves

below that edge will become extinct: $p_S(i) = 1 - \prod_{j \in C_i}(1 - a_j)$.

Using these survival probabilities, for a given $\vec{a}$, the integrand in Equation 8.3 can be found by summation over the edges in the tree:

$$\begin{aligned} \mathbb{E}(PD|S,\vec{a}) - \mathbb{E}(PD|\emptyset,\vec{a}) &= \sum_i \lambda_i (p_S(i) - p_\emptyset(i)) \\ &= \sum_{i:C_i \cap S \neq \emptyset} \lambda_i (1 - p_\emptyset(i)) \\ &= \sum_{i:C_i \cap S \neq \emptyset} \lambda_i \prod_{j \in C_i} (1 - a_j) \end{aligned} \tag{8.4}$$

as $p_S(i) = p_\emptyset(i)$ unless S contains a leaf below edge i. Substitution of Equation 8.4 in Equation 8.3 yields:

$$\begin{aligned} \mathbb{E}_{\vec{a}}(\mathbb{E}(PD|S,\vec{a}) - \mathbb{E}(PD|\emptyset,\vec{a})) &= \sum_{i:C_i \cap S \neq \emptyset} \lambda_i \int_{\vec{a}} \phi(\vec{a}) \prod_{j \in C_i} (1 - a_j) d\vec{a} \\ &= \sum_{i:C_i \cap S \neq \emptyset} \lambda'_i. \end{aligned} \tag{8.5}$$

This is simply the phylogenetic diversity of a set S using the tree with modified edge lengths λ'_i. Optimal solutions to the uNAP are therefore simply the solutions to the NAP on this modified tree.

8.3.1 Alternative Objective Functions

The objective function for the uNAP seems to be the obvious extension of the NAP. However this introduces a bias towards low survival probabilities where higher increases can be obtained as more species will become extinct in the absence of conservation. Whether this is appropriate depends on a value judgement, do we want our solution to protect a greater proportion of the 'at risk' PD when more is at stake or should the same proportion of 'at risk' PD be protected across the range of plausible survival probabilities? We now introduce two alternative objective functions that achieve the latter objective. Firstly, denoting the set of all species by X, we have the proportion of the 'at risk' PD that is conserved by S:

$$\mathbb{E}_{\vec{a},\vec{b}}\left(\frac{\mathbb{E}(PD|S,\vec{a},\vec{b})-\mathbb{E}(PD|\emptyset,\vec{a},\vec{b})}{\mathbb{E}(PD|X,\vec{a},\vec{b})-\mathbb{E}(PD|\emptyset,\vec{a},\vec{b})}\right). \tag{8.6}$$

Secondly we have the proportion of the maximum attainable PD increase that is achieved by S:

$$\mathbb{E}_{\vec{a},\vec{b}}\left(\frac{\mathbb{E}(PD|S,\vec{a},\vec{b})-\mathbb{E}(PD|\emptyset,\vec{a},\vec{b})}{\mathbb{E}(PD|\hat{S},\vec{a},\vec{b})-\mathbb{E}(PD|\emptyset,\vec{a},\vec{b})}\right). \tag{8.7}$$

where $\hat{S}$ is an optimal solution to the NAP for each value of $\vec{a}$. Both of these alternative objective functions become equivalent to the NAP objective function if survival probabilities are known with certainty.

To utilise these objective functions for the restricted scenario described in Theorem 15, an alternative branch rescaling is performed:

$$\lambda' = \lambda_i \int_{\vec{a}} \frac{\phi(\vec{a})}{d(\vec{a})} \prod_{j\in C_i}(1-a_j)d\vec{a},$$

where $d(\vec{a})$ is the denominator in Equations 8.6 and 8.7 (respectively $PD(X)-PD(\emptyset)$ or $\hat{PD}$). Note that this complicates the uNAP as $d(\vec{a})$ depends on $\vec{a}$ and in the second case calculating $\hat{PD}$ requires a solution to the normal NAP to be obtained for each value of $\vec{a}$.

8.3.2 Computational aspects of rescaling branch lengths

As we have shown, solutions to the uNAP are obtainable by applying conventional methods for solving the NAP to a tree with rescaled branch lengths (Equation 8.2). Depending on the complexity of the survival probability density, Equation 8.2 may possess analytic solutions or require numerical solution. Here we provide an simple solution to Equation 8.2 for independent uniform survival probability densities, $\phi_j(a_j)$. By assuming that the survival probability densities are independent we have:

$$\begin{aligned}\lambda' &= \lambda_i \int_{\vec{a}} \phi(\vec{a}) \prod_{j \in C_i} (1 - a_j) d\vec{a} \\ &= \lambda_i \prod_{j \in C_i} \int_{a_j=0}^{1} \phi_j(a_j)(1 - a_j) da_j. \end{aligned} \tag{8.8}$$

Consider survival probability densities that are uniform on some interval $[l_j \; u_j]$:

$$\phi_j(a_j) = \begin{cases} \frac{1}{u_j - l_j} & \text{if } l_j \leq a_j \leq u_j \\ 0 & \text{otherwise.} \end{cases}$$

Substitution of these survival probability densities in Equation 8.8 yields:

$$\begin{aligned}\lambda' &= \lambda_i \prod_{j \in C_i} \int_{l_j}^{u_j} \frac{1}{u_j - l_j} (1 - a_j) da_j \\ &= \lambda_i \prod_{j \in C_i} \left(1 - \frac{u_j + l_j}{2}\right) \end{aligned} \tag{8.9}$$

8.4 Applications

In this section we apply the uNAP to two examples. First we consider a single parameter scenario using simulated trees. This scenario shows that point estimates provide solutions that are robust to uncertainty in the survival probabilities unless zero value point estimates are used. Secondly we apply this methodology to the Madagascan Lemurs. Utilising the threat status from IUCN (2007) each of the lemurs is assigned a survival probability. There is much uncertainty in the conversion from IUCN threat status categories to survival probabilities, here we show that the species prioritisations are robust to this uncertainty.

8.4.1 Single parameter scenario

We begin by considering a simple NAP where all species have the same survival probability if they are not conserved ($a_j = a$), they survive with

certainty if they are conserved ($b_j = 1$) and they all cost the same to conserve ($c_j = c$). This last restriction ensures that we can simplify the problem to one of finding the optimal set of k species to conserve (where k is determined by the budget, see Theorem 10). In this application we consider uncertainty in the conservation timescale – the magnitude of the survival probabilities, a.

First we use the conventional NAP to obtain solutions for given values of a on one hundred sampled Yule trees (see chapter 3 and Hartmann et al. (a)). For each tree and value of k, optimal sets were found using the algorithms in chapter 7 for a range of values of a.

If the true value of a differs from the value used to find a set, that set may no longer be optimal. Figure 8.2 shows how the quality of a set decreases as the true value of a changes from that used to obtain the set. Note that solutions obtained with a zero value of a are not robust across survival probabilities. As noted in Redding et al. (2008) this is due to the ultrametricity of the tree (the distance between the root and each leaf is fixed in a Yule tree) which ensures that for each species there are a set of solutions that include that species. Accordingly whilst the set of optimal solutions corresponding to $a = 0$ may contain some solutions that perform well across a range of survival probabilities it will also contain many solutions that perform poorly.

We now use the uNAP to obtain solutions that are robust across a range of survival probabilities. Since the survival probabilities are all equal in this scenario ($a_j = a$) it suffices to use a probability density on one parameter, a. We consider uniform probabilities over different ranges as illustrated in Figure 8.2; uniform probability densities were arbitrarily chosen for their simplicity. Surprisingly there is only a minimal difference in the robustness of each NAP solutions with a survival probability a and the uNAP solution with a uniform density centered on a. For practical applications this suggests that the additional complexity introduced by the uNAP may be insufficient to warrant its usage over the simpler NAP.

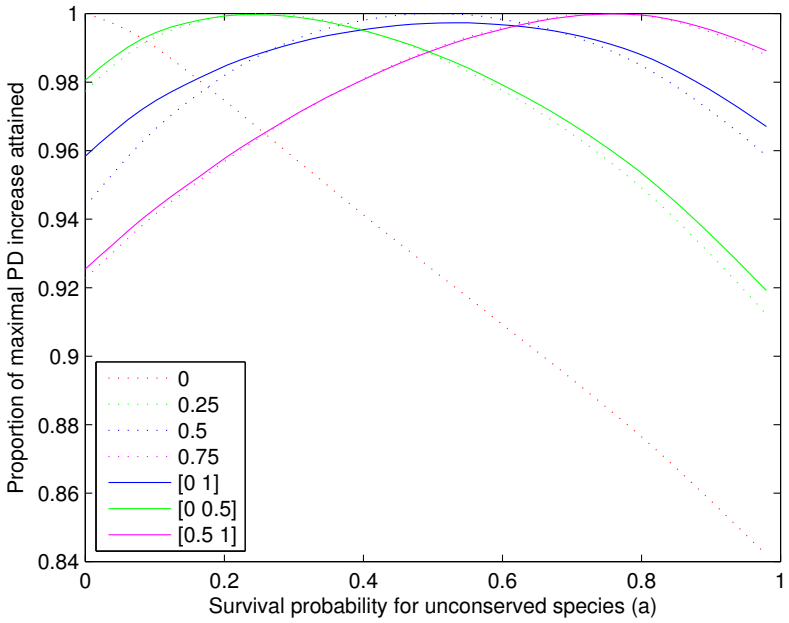


Figure 8.2: The x axis shows the probability assigned to the survival of all unconserved species. The y axis shows the proportion of the maximal PD increase that is obtained for a particular solution. For example consider the red dotted line. If the survival probability of the species is 0.6 the solutions corresponding to this line are expected to produce an increase in PD that is about 91% of the maximum increase possible at that survival probability. The lines depicted in this figure correspond to sets of 5 species being conserved from a set of 20 on one hundred randomly generated Yule trees. The dotted lines correspond to solutions obtained using point estimates for the survival probabilities. Notably solutions obtained using a point estimate of 0 are more sensitive to changes in the survival probabilities than those corresponding to other point estimates. The solid lines correspond to solutions produced using the uNAP and a uniform probability density over the indicated ranges. The similarity of these solutions with those obtained using the NAP and point estimates in the middle of the corresponding ranges is remarkable.

8.4.2 Application to Madagascar's Lemurs

In this section we consider species prioritisations for the Madagascan lemurs. There are currently 62 recognised species of lemur and a phylogeny for these species has recently been constructed by Sebastien Rioux-Paquette and Arne Mooers in Hartmann et al. (b) (see Figure 8.3C). The extinction threat of 50 of these species has been classified in IUCN (2007). Of the 62 lemurs, 9 are critically endangered, 17 are endangered and 18 are vulnerable. Clearly without intervention some extinctions will occur in the near future. Furthermore Madagascar has been recognised by the 'WWF for Nature' as one of their 19 global priority areas and lemurs have been identified as a priority species for their conservation work.

Here we apply the NAP and uNAP to the lemur tree. As before we assume that conserved species will survive with certainty and all species cost the same amount to conserve. We will show that our solutions are robust to highly uncertain survival probability estimates and that a significant biodiversity gain can be obtained by using the phylogeny and the extinction risks. We do not suggest that the lemur prioritisations we present are final – other factors excluded from this analysis (such as conservation costs) would need to be incorporated to produce a final realistic prioritisation.

The IUCN threat status is categorical with five levels of extinction risk, to apply the NAP it is necessary to convert these categories to probabilities. Mooers et al. summarises different relationships between the IUCN extinction risk categories and extinction probabilities, as outlined in Table 8.1. These relationships differ greatly, thereby presenting an important question: how robust are solutions to the choice of relationship and should the uNAP (Theorem 15) be used to incorporate our uncertainty about the relationships?

Survival probabilities for the lemurs were obtained by converting the IUCN categories to probabilities using both the IUCN and Isaac relationships (Table 8.1). Using these probabilities and Equation 8.2 the lemur tree (Figure 8.3C) was rescaled, producing the trees in Figure 8.3A and B. The greedy algorithm (Theorem 7) can be applied to these rescaled trees to find optimal sets of species to conserve. Table 8.2 shows the order in which species are selected and the cumulative PD represented by those species. The PD

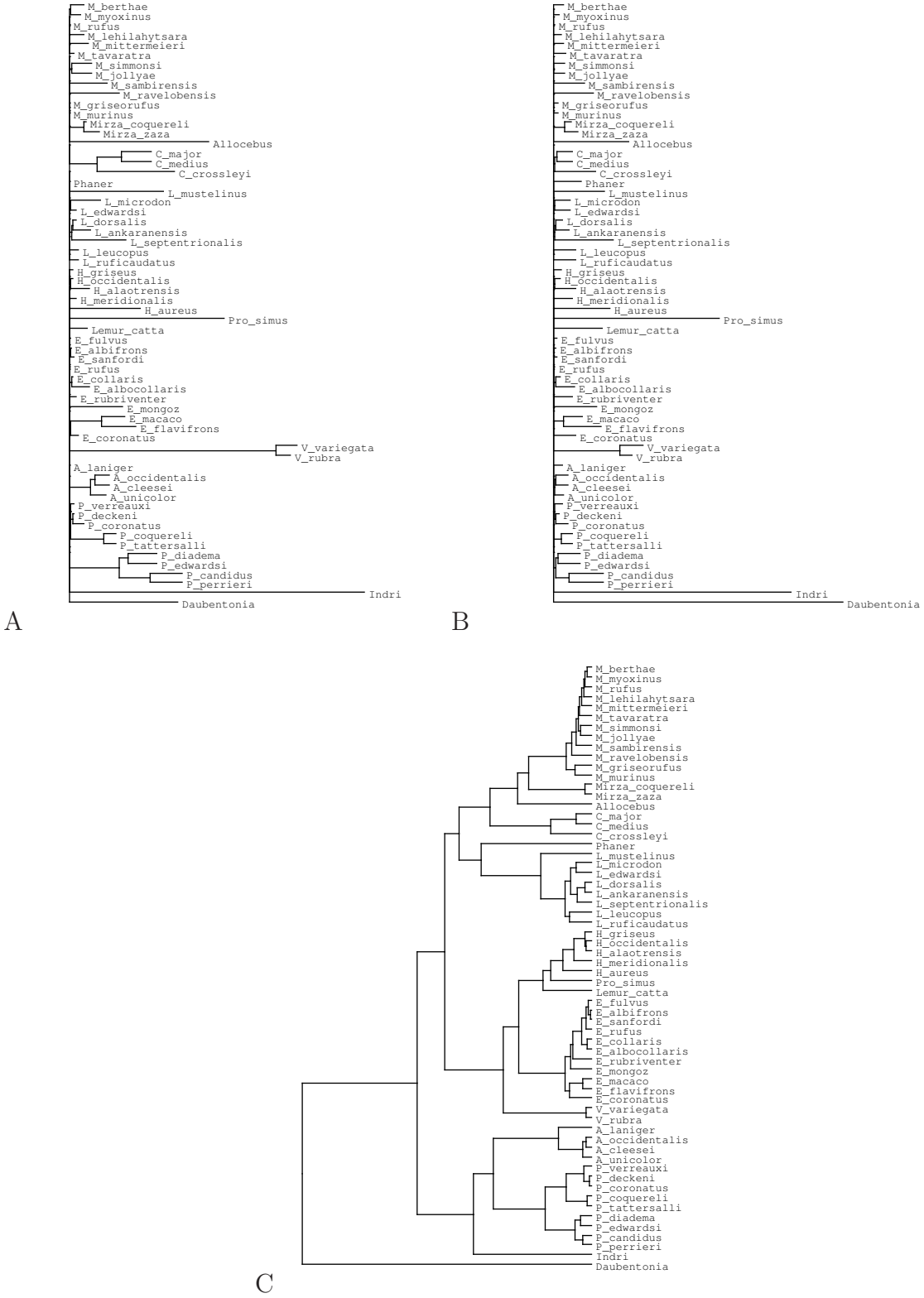


Figure 8.3: The phylogenetic tree for the Madagascan lemurs from Hartmann et al. (b). C: The unscaled tree. A and B: The tree rescaled using the survival probabilities in Table 8.1. The greedy algorithm an be applied to these trees to provide optimal solutions corresponding to these survival probabilities.

Category	IUCN	Isaac
Critical (CR)	0.001	0.6
Endangered (EN)	0.33	0.8
Vulnerable (VU)	0.9	0.9
Least concern (LC)	1	0.97
Data deficient (DD)	0.5	0.9

Table 8.1: In Mooers et al. a summary of possible survival probabilities for each IUCN category were given. Two of the most accepted sets of probabilities are those from Mace and Lande (1991) and Isaac et al. (2007). Following Mooers et al. we will respectively refer to these probabilities as the IUCN and Isaac probabilities.

has been expressed in terms of the proportion of the rescaled edge lengths and therefore corresponds to the proportion of 'at risk' PD that would be secured by conserving each set of species.

An important question is how good these species prioritisations perform if the survival probabilities with which they were produced are incorrect. To assess this we will consider the performance of the species prioritisations for two scenarios A) The IUCN survival probabilities are correct and B) the Isaac survival probabilities are correct. In Figure 8.4A and B the proportion of at risk PD represented by various sets of species is shown. Consider Figure 8.4A, the set obtained by applying the greedy algorithm to the tree rescaled using the IUCN probabilities (Figure 8.3A) produces the optimal solution. Using the rescaled tree based on the Isaac probabilities (Figure 8.3B) results in a slightly inferior solution. Both of these solutions are significant improvements over a set obtained by randomly selecting species or by maximising PD on the original tree (Figure 8.3C).

If we are uncertain whether the IUCN or Isaac probabilities are most appropriate then we may wish to find a set that compromises between these two extremes. A very simple possibility (i) is to take the mean survival probability for each species. An alternative (ii) is to use uNAP. This latter approach was applied, using a uniform distribution between the two extreme sets of probabilities (IUCN and Isaac). Figure 8.5 shows the effect of using these methods. Overall if each scenario is equally likely both of the compromising solutions have a higher expected PD than a solution produced by assum-

	IUCN			Isaac		
	Species	Threat	Cum. PD	Species	Threat	Cum. PD
1	Indri	EN	0.13	Daubentonia	VU	0.15
2	V_variegata	CR	0.23	Indri	EN	0.27
3	Pro_simus	CR	0.3	Pro_simus	CR	0.36
4	Allocebus	DD	0.36	V_variegata	CR	0.41
5	P_candidus	CR	0.41	Allocebus	DD	0.44
6	Daubentonia	VU	0.46	L_septentrionalis	CR	0.48
7	C_crossleyi	DD	0.51	H_aureus	EN	0.5
8	L_mustelinus	DD	0.55	L_mustelinus	DD	0.53
9	H_aureus	EN	0.58	P_candidus	CR	0.56
10	E_flavifrons	CR	0.61	Lemur_catta	VU	0.58
11	L_septentrionalis	CR	0.64	E_flavifrons	CR	0.61
12	E_mongoz	EN	0.66	E_mongoz	EN	0.63
13	C_major	DD	0.69	C_crossleyi	DD	0.65
14	M_ravelobensis	EN	0.71	M_ravelobensis	EN	0.67
15	P_coquereli	EN	0.73	P_perrieri	CR	0.69
16	A_occidentalis	EN	0.75	M_sambirensis	EN	0.7
17	M_sambirensis	EN	0.76	Phaner	LC	0.72
18	P_diadema	EN	0.78	P_diadema	EN	0.73
19	P_perrieri	CR	0.8	Mirza_zaza	EN	0.74
20	L_microdon	DD	0.81	P_edwardsi	EN	0.76

Table 8.2: The top twenty lemur species listed in order of importance. The two lists correspond to the different sets of survival probabilities listed in Table 8.1. The cumulative proportion of the 'at risk' PD represented for the two sets of survival probabilities is also listed.

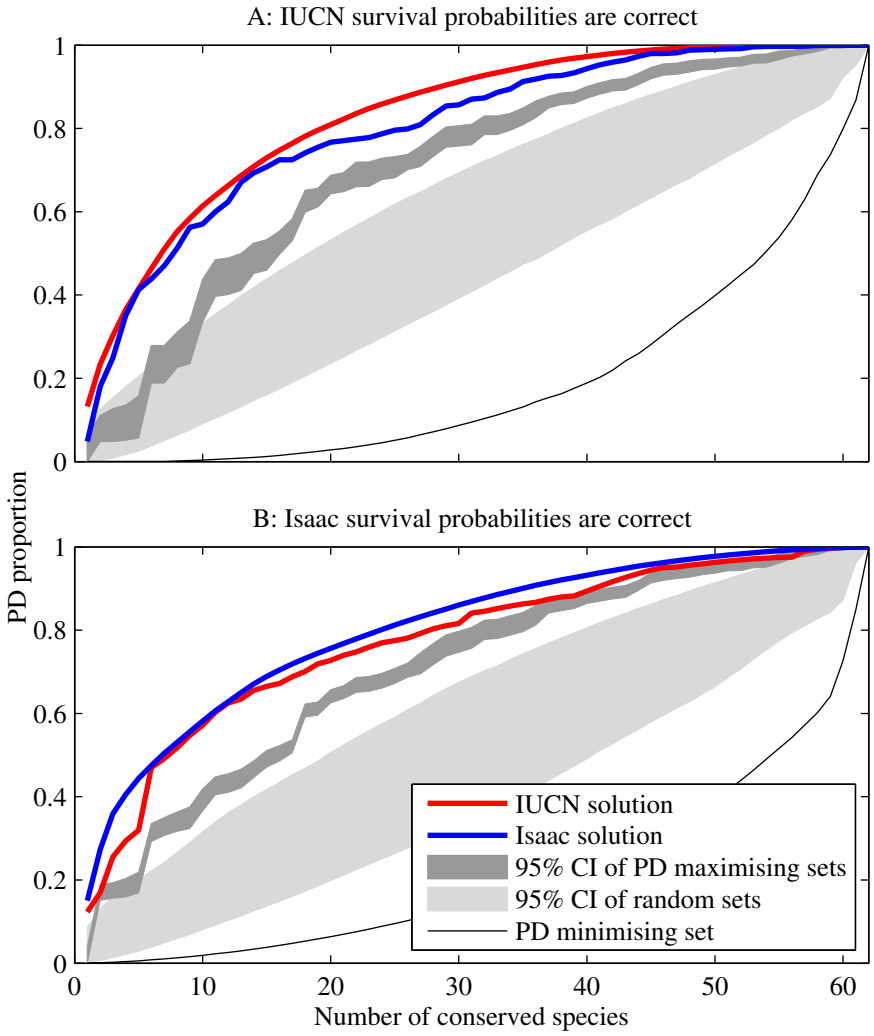


Figure 8.4: Two scenarios are considered, A: the IUCN survival probabilities are correct, B: the Isaac survival probabilities are correct. The 'at risk' PD represented by sets of various sizes obtained using an assumption of both sets of survival probabilities is depicted. A 95% CI for random sets of species and PD maximising sets that do not take into account survival probabilities are shaded in gray.

ing the IUCN or Isaac probabilities. The difference between solutions (i) and (ii) is minimal with (i) performing slightly better if the IUCN survival probabilities are actually correct.

8.5 Concluding comments

In this chapter we have considered the effect of uncertain species survival probabilities on solutions produced by the Noah's Ark Problem. We have provided an extension to the NAP, the uNAP, which allows uncertainty about species survival probabilities to be included in this framework. The uNAP permits arbitrarily complicated probability densities to be used for the species survival probabilities. These densities may be as simple as a uniform density between zero and one (reflecting complete uncertainty about the survival probability of a species) and it is expected that in many applications relatively simple densities covering a large range of survival probabilities would be used.

The single parameter scenario we investigated considered the case where all survival probabilities were equal but their value was not known with certainty. This corresponds to the situation where there is uncertainty about the magnitude of survival probabilities or about the appropriate time scale for conservation management. We considered this scenario using simulated Yule trees. There was only a small advantage in using the uNAP over the standard NAP. This indicates that the solutions to the standard NAP are robust to uncertainty in the survival probabilities. The exception to this was where a survival probability of zero was used, solutions obtained using this estimate performed poorly if the true survival probabilities were non-zero. This is an extremely important point as studies that aim to simply maximise PD implicitly assume a zero survival probability.

We applied the NAP to the Madagascan lemurs, using a recent phylogenetic tree and the threat status from IUCN (2007). The solutions obtained were robust to the highly uncertain relationship between the species threat categories and survival probabilities. Our extension to the NAP provided only a small overall gain in the expected 'at risk' PD that was represented. This again suggests that approximate survival probabilities are sufficient for

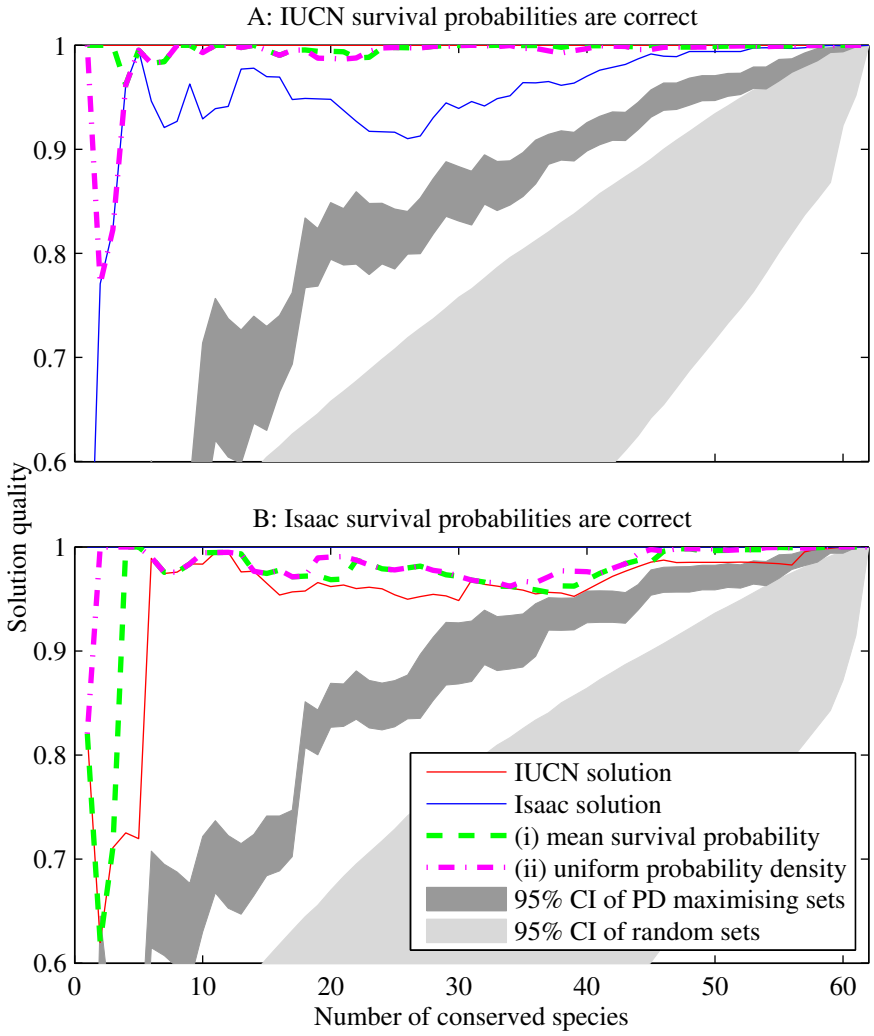


Figure 8.5: The same two scenarios shown in Figure 8.4 are considered. Here we rescale the represented PD such that a value of one corresponds to the optimal solution. Sets produced by assuming a mean survival probability and uniform range of survival probabilities between these two extremes are shown in addition to the sets in Figure 8.4. These methods provide a good compromise if the true survival probabilities are unknown.

obtaining good solutions. Lastly for this example, inclusion of both a phylogenetic tree and survival probabilities produced significant increases in the amount of 'at risk' PD that was represented.

In this chapter we have considered only one source of uncertainty in the NAP framework. The other sources are the survival probabilities of conserved species, the conservation costs and the phylogenetic tree. Arguably the least important of these are the conservation costs as they can in some sense be regulated. Including uncertainty for the survival probability of conserved species is an important problem and one that promises to be computationally complex.

Another important source of uncertainty is the phylogenetic tree itself. If each edge length has a probability distribution associated with it, the mean of this distribution can simply be used. Uncertainty in the tree shape is a far more complicated problem, on which interesting work is currently being done (Spillner et al.; Minh et al.). Hopefully NAP solutions will be robust to this source of uncertainty and it will suffice to use the 'best' tree.

Chapter IX

When should phylogenies guide conservation?

In previous chapters we have seen that including phylogenies in conservation management is a complex task. This complexity may reduce the ability of a conservation manager to communicate their decisions to stakeholders and may not fit within limitations of existing management frameworks (Redding et al., 2008). Given these problems the importance of including phylogenetic information should be carefully assessed. In this chapter we introduce two upper bounds on the possible benefits of including a phylogeny in conservation management – the expected value of perfect choice ($EVPC$) and the maximum value of perfect choice ($MVPC$). For a given conservation problem the value of these measures can be used to evaluate the importance of the phylogenetic information before deciding whether it should be included.

In many cases a phylogeny is not readily available and may take extra resources to obtain, possibly reducing the resources available for primary conservation work. Consequently it is desirable to obtain an estimate of the possible benefit of including a phylogeny before one is constructed. To assist in this decision we investigate the distribution of $EVPC$ and $MVPC$ for phylogenies produced by evolutionary models. Some limited knowledge of a phylogeny (eg. expert opinion about its tree balance) may be available before it is constructed, so we highlight some characteristics of phylogenies that effect their importance for conservation.

9.1 The value of perfect choice

The measures we introduce are in the spirit of Carl Walters' expected value of perfect information ($EVPI$; Walters (1986)). They give an upper bound

on the expected and maximum benefit that inclusion of PD can provide. The first measure is the expected value of perfect choice ($EVPC$). $EVPC$ gives the biodiversity increase obtained when the best set of species of a given size is selected over a random set of species of that size:

$$\begin{aligned} EVPC &= PD(\text{best set with k species}) - E(PD(\text{any set with k species})) \\ &= \max_S PD(S) - \mathbb{E}_S[PD(S)], \end{aligned}$$

where S ranges over all subsets of k species. If $EVPC$ is low, the phylogenetic tree may be unimportant for conservation. If $EVPC$ is high then including the phylogeny may result in a much greater representation of biodiversity.

$EVPC$ is an appropriate bound if omitting PD from conservation management will result in random species selections from the perspective of PD. However some conservation methods will select species non-randomly from a phylogenetic perspective. In this case the expected biodiversity gain may exceed $EVPC$. An example of this is when a group of charismatic or commercially valuable, but closely related species would be selected. A set containing these closely related species will have lower PD than a random set of species, consequently the possible gain of using a phylogeny for conservation management may exceed $EVPC$.

The second measure is the maximum value of perfect choice ($MVPC$) which provides an absolute upper bound on the attainable biodiversity gain. This measure gives the biodiversity increase obtained when the best set of species is selected instead of the worst set of species:

$$\begin{aligned} MVPC &= PD(\text{best set with k species}) - PD(\text{worst set with k species}) \\ &= \max_S PD(S) - \min_S PD(S). \end{aligned}$$

$EVPC$ and $MVPC$ are upper bounds on the attainable biodiversity gain. These bounds are only attainable if the best set of species can be chosen under any additional constraints that may be present on the conservation problem. Consider an optimal solution to a conservation problem without phylogenetic information. Due to the constraints discussed in chapter 7, such as different conservation costs, it may not be possible to swap a species that

contributes little to the biodiversity for one that does. Instead it may be necessary to swap several species that contribute little to biodiversity for a single species that makes a greater contribution but costs more to conserve. Such constraints will mean that *EVPC* and *MVPC* are rarely attainable. *EVPC* (although not under that name) has previously been considered in studies exploring biodiversity loss under random extinction (Nee and May, 1997; Heard and Mooers, 2000).

9.2 Application to the Yule model

Using evolutionary models it is possible to investigate the behaviour of *EVPC* and *MVPC* across a large range of possible trees. We consider the Yule model (Yule, 1924) extensively, beginning with an analytic method for calculating the expectation of *EVPC* for Yule trees.

9.2.1 Analytic expectation of EVPC

We denote the expected value of $EVPC(\mathcal{T})$ for n-species Yule trees by:

$$\mathbb{E}_{\mathcal{T}}\left[EVPC(\mathcal{T})\right] = \mathbb{E}_{\mathcal{T}}\left[\max_{S} PD(S)\right] - \mathbb{E}_{\mathcal{T}}\left[\mathbb{E}_{S}\left[PD(S)\right]\right],$$

where the expectation is over all trees with n species.

The maximum PD of a set of size k depends only on the timing of the first k speciation events, not on the tree shape. Until the kth speciation event all edges will be spanned by an optimal set, after this point only k edges extant at a given time may be conserved. Utilising the notation from section 2.3.2 we have:

$$\mathbb{E}_{\mathcal{T}}\left[\max_{S} PD(S)\right] = \mathbb{E}_{\mathcal{T}}\left[\sum_{i=1}^{k} i(\Lambda_{i-1} - \Lambda_i) + k(\Lambda_k)\right],$$

where the first term corresponds to the time before the kth speciation event where all edges are spanned by an optimal set and the second term corresponds to the time thereafter where only k lineages are spanned. This simplifies to give:

$$\mathbb{E}_{\mathcal{T}}\left[\max_{S} PD(S)\right] == t + \sum_{i=1}^{k-1} \mathbb{E}\left[\Lambda_i | n, t\right],$$

The expectation of the speciation events is given in Theorem 2, it is therefore straightforward to find the expected maximum PD.

The expected PD over all sets of species depends on the tree shape as well as the timing of the speciation events. For a given tree shape, τ, the probability with which an edge will be spanned by a random set depends only on the number of descendants it has. Recalling that C_i is the number of descendants of an edge i we obtain:

$$\mathbb{E}_{\mathcal{T}}[\mathbb{E}_S[PD(S)]] = \sum_{\tau} p(\tau|n) \sum_{e} \mathbb{E}[\lambda_e|\tau] \left(1 - \frac{\binom{n-C_e}{k}}{\binom{n}{k}}\right).$$

The term being summed is the the expected length of an edge ($\mathbb{E}[\lambda_e|\tau]$) times the probability that it is spanned by a random set of species. This is summed over all edges and all possible tree shapes (which are weighted by their probability of occurring). This can be calculated using either of the methods presented in chapter 2.

9.2.2 Characteristics of EVPC *and* MVPC

To calculate the expected value of $MVPC$ it is necessary to find a PD minimising set. To find this set the dynamic programming algorithm described in chapter 6 was used, consequently calculating the $MVPC$ distribution analytically is not straightforward. The results in this section were therefore produced using sampled trees obtained with the algorithms described in chapter 3.

We considered conserving species from twenty-species trees and sampled 200 trees. Figure 9.1 shows the expected PD of the PD maximising set, minimising set, and of random sets. The expected $EVPC$ and $MVPC$ are readily calculated from these curves. Figure 9.1 illustrates that in absolute terms $EVPC$ and $MVPC$ are highest when an intermediate number of species are being selected. Figure 9.2 shows the distribution of $EVPC$ and $MVPC$, the spread of this distribution is high, this variability suggests that $EVPC$

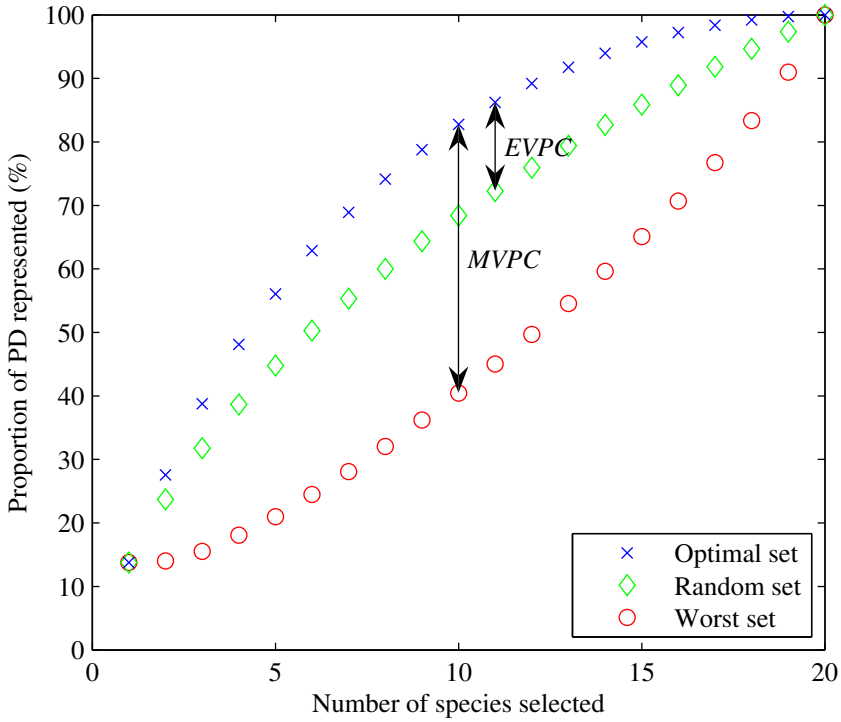


Figure 9.1: This figure depicts the expected proportion of the total PD in a tree that is represented (y axis) by picking sets of species of different sizes (k; x axis) from twenty-species Yule trees. The best, worst and random sets were considered. The expected value of $EVPC$ is the difference between the expected value of the best set and a random set. For $MVPC$ it is the difference between the expected value of the best set and the worst set.

and $MVPC$ will need to be evaluated on a case by case basis.

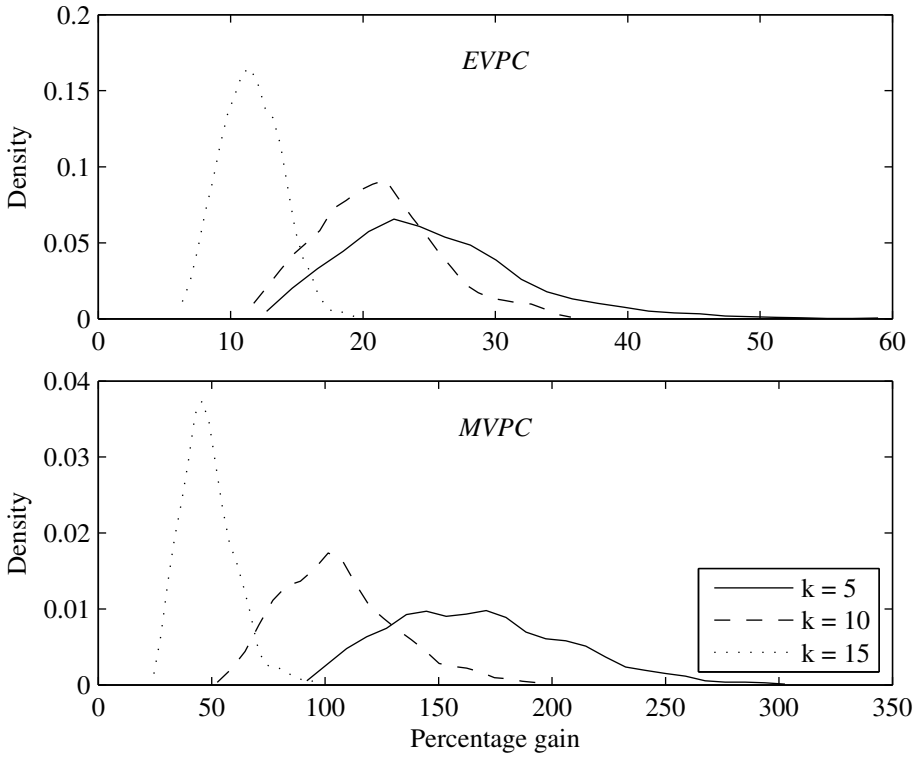


Figure 9.2: This figure shows the distribution of *EVPC* and *MVPC* for the situation depicted in figure 9.1. Set sizes (k) of five, ten and fifteen species were considered.

9.3 Application to 'real' trees

A set of trees was obtained from McPeek and Brown (2007). These trees are derived from real data for a broad range of organisms at a species level, hence they correspond to a range of potential biodiversity conservation problems. Many of the trees in the sample were not ultrametric, some of these featured a small number of highly diverged species such as the tree in Figure 9.3. To remove the effect due to such species, only trees that were nearly ultrametric were considered. We defined near ultrametric trees as those where the distance between the root and the leaves differed by less than 3%. For brevity we subsequently refer to this sample of 46 trees as the 'real' trees.

EVPC and *MVPC* were calculated for the 'real' trees. These points are depicted in Figure 9.4, along with their mean and 95% confidence interval. The sizes of the trees varied, hence the values for the proportion of species conserved differ between trees. Consequently these values were binned and the mean and 95% confidence intervals were calculated for each bin. Two trees with extreme (high and low) $EVPC/MVPC$ values are highlighted in Figure 9.4 and depicted in Figure 9.5.

Figure 9.3: A non-ultrametric tree with a single highly diverged species. *EVPC* and *MVPC* will be extremely high for small set sizes (small k) as it is imperative that the diverged species is selected. *EVPC* and *MVPC* values for this tree are shown in Figure 9.4.

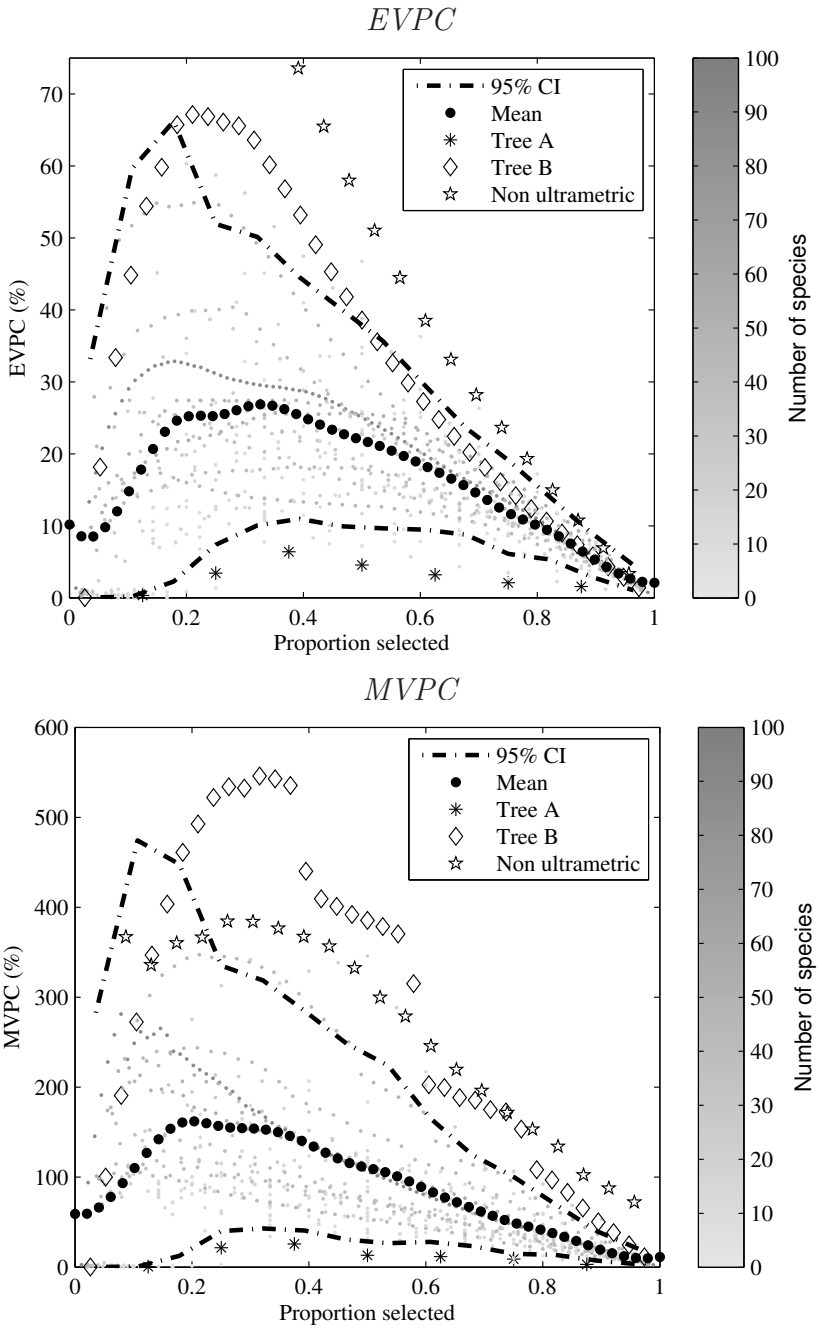


Figure 9.4: This figure shows *EVPC* (top panel) and *MVPC* (bottom panel) for the sample of 'real' trees. Points for individual trees are plotted and colour coded according to the tree's size. The points corresponding to two trees with extreme values (see Figure 9.5) are highlighted. A smoothed mean of *EVPC*/*MVPC* is depicted along with a 95% confidence interval. The points corresponding to the non-ultrametric tree in Figure 9.3 are also shown (all other points and lines are for near-ultrametric trees).

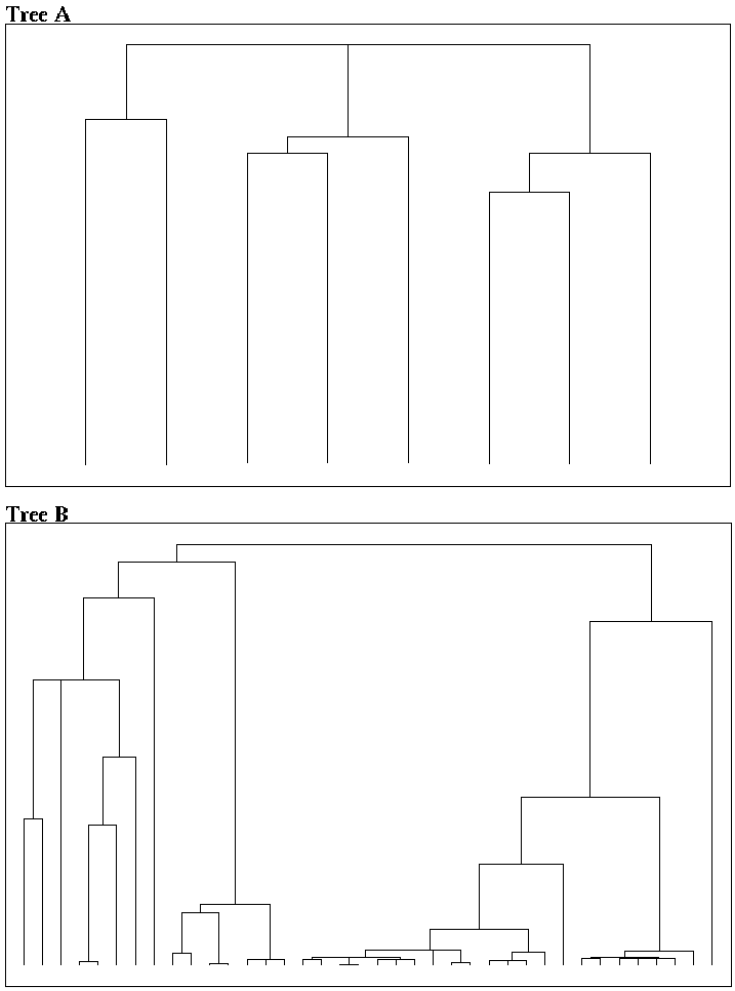


Figure 9.5: This figure shows the two trees indicated in Figure 9.4. Tree A has low $EVPC/MVPC$ whereas Tree B has high $EVPC/MVPC$.

9.4 Tree characteristics influencing EVPC

To calculate *EVPC* or *MVPC* for a given conservation problem the corresponding phylogenetic tree must be available. In many cases much effort and substantial resources will be required to construct a useful phylogeny. In this chapter we have seen that the value of a constructed phylogeny for conservation management varies greatly. Often some limited information may be available about the phylogeny, eg. it may be known that many species are quite young or that the tree is expected to be quite imbalanced (as discussed in chapter 4). This information may be useful to determine whether a constructed phylogeny is expected to be sufficiently useful for conservation management to allocate resources to its construction.

Two attributes about which we may have some information are tree imbalance and the relative age of species. Here we consider the correlation between indices that summarise these characteristics and *EVPC*. Tree imbalance statistics were introduced in chapter 4, here we consider only I_C as these two indices are highly correlated and I_C is more widely used. Pybus and Harvey (2000) introduced a γ statistic which gives an indication of the timing of the speciation events in a tree. For Yule trees the γ statistic is normally distributed with a zero mean. Negative values indicate that speciation events in a tree occurred closer to the root than expected and positive values indicate that speciation events occurred closer to the tips (species are relatively young).

We denote the time between the speciation event that created the ith and the $i+1$th species by σ_i and ignore the length of the root edge. Using this notation the γ index of a n-species tree is given by:

$$\gamma = \frac{\frac{1}{n-2}\sum_{i=2}^{n-1}\left(\sum_{k=2}^{i} k\sigma_k\right) - \frac{K}{2}}{K\sqrt{\frac{1}{12(n-2)}}}, \qquad K = \sum_{j=2}^{n} j\sigma_j$$

As discussed in chapter 4, for memoryless models the tree shape is independent of the distribution of speciation events. Consequently the imbalance statistics and the γ statistic measure independent aspects of phylogenetic trees produced by such processes. A linear regression was performed be-

Variables	Tree Source		
	Real trees	Birth-death process	
Constant	13.5 ± 3.5	10.5 ± 0.6	10.5 ± 0.5
Tree size	0.32 ± 0.08	0.33 ± 0.02	0.32 ± 0.02
γ	2.8 ± 0.6	-0.16 ± 0.06	0.05 ± 0.07
I_c	12 ± 6	13.5 ± 0.5	13.5 ± 0.5
Extinction rate			-0.9 ± 0.3
r^2	0.55491	0.13773	0.13947

Table 9.1: This table contains regression coefficients with 95% confidence intervals from a linear regression of *EVPC* and the indicated explanatory variables. Both trees produced by a constant rate birth-death process and 'real' trees as described in the main text were considered. In all cases half the species in a given tree were selected. *EVPC* was expressed as percentage gain, so for example the expected *EVPC* of a real twenty-species tree with $\gamma = 0.5$ and $I_c = 0.5$ is $13.5 + .32 \times 20 + 2.8 \times 0.5 + 12 \times 0.5 = 27$; including a phylogeny in conservation management is expected to bring a gain of 27%.

tween *EVPC* and various tree statistics and the extinction rate for trees produced by the constant rate birth-death model. This regression was performed for both the sample of 'real' trees and trees produced by the constant rate birth-death process with a uniform distribution of extinction rates. The coefficients calculated for the regression are presented in Table 9.1.

Two models were fitted to the sample from the birth and death process, one with the extinction rate as a parameter and one without. There is a high degree of correlation between γ and the extinction rate, the model without the extinction rate permits a parameter comparison with the regression for the 'real' trees. All the models fitted were highly significant, however the model explains only a small amount of the variation in *EVPC* (as seen from the low r^2 values) for the Yule sample.

For the sample of 'real' trees there is a pronounced link between γ and *EVPC*, trees with speciation events occurring closer to the present are likely to have higher *EVPC*. For the trees sampled from the constant rate birth-death process this relationship is significant but weak and reversed. The strongest relationship is with the imbalance index I_c; for both samples, more imbalanced trees have greater *EVPC*.

Figure 9.5 shows trees with extremely high or low *EVPC*. The 'real' tree in Figure 9.5B has extremely high *EVPC* as there are many very young species and a few very old species, this is an example of a tree with a high γ value that is unlikely to occur under the constant rate birth-death model. Trees like this are responsible for the high correlation between γ and *EVPC* for 'real' trees. If γ is excluded from the models, the r^2 value for both tree samples is much closer (0.17 and 0.14), this indicates that the strange split distribution in speciation times (some early speciations and many late speciations) explains much of the variation in *EVPC* for real trees. An investigation of the origin of this split distribution in the 'real' trees is beyond the scope of this work but would be an interesting future avenue of enquiry.

9.5 Concluding comments

The two measures introduced in this chapter – *EVPC* and *MVPC* – place an upper bound on the benefit obtainable by including phylogenetic information in a conservation decision. Which of these bounds is most appropriate will depend on the nature of the solutions produced in the absence of phylogenetic information. If it is suspected that these solutions would have lower PD than a random sample, *MVPC* may be a more appropriate measure.

For both the sample of 'real' trees and trees produced by the constant rate birth-death process, expected *EVPC* values were typically 20% and *MVPC* values were between 100% and 150% when half the species were conserved. The variation in these values is large with the 95% confidence interval being approximately ±50% of the expected *EVPC*/*MVPC*. The expected gain for a given situation may therefore vary between unimportant (eg. 10%) and highly important (eg. 30%).

We examined the correlation of *EVPC* with various tree / problem characteristics in an attempt to provide a estimate of the expected gain in situations where little is known about the phylogeny. For the trees examined there was a high correlation with tree imbalance – more imbalanced trees are expected to have a higher *EVPC*. This is because more imbalanced trees are likely to have some basal species with long pendant edges that are crucial to identify. For the 'real' trees considered there was also a strong correlation

with the γ statistic. This was due to the split distribution of speciation times observed in some 'real' trees (eg. Figure 9.5B), an aspect worthy of further investigation.

In this chapter we have seen that the value of phylogenetic information can be much higher if management is implicitly including phylogenetic information (as shown by the large difference between $EVPC$ and $MVPC$). One possible mechanism would be over-represented clades due to the charismatic or commercial value of a group of closely related species. This suggests that the real gain is closer to the higher $MVPC$, however such situations may be those with the least flexibility for selecting different sets of species.

$EVPC$ and $MVPC$ are upper bounds, the real value is likely to be lower. How much lower will depend on the penalties other management restrictions place on choosing different sets of species. For example if there is a predefined set of species of great economic or charismatic value, the penalty of not choosing this set of species may be so great (politically or financially) that phylogenetics simply becomes irrelevant.

On a concluding note, Nee and May (1997) noted that under their model for clades of 50 and 500 species about half the PD was preserved by saving 20% of the species. Unfortunately if extinction/conservation choice is non-random our study suggests that as little as 20% of PD may be represented.

Chapter X

Species specific indices

Many indices for measuring or ranking the distinctiveness of a single taxon have been proposed. These indices have the advantage of being easy to compute and, as each taxon is assigned some value, they can easily be included in a complex decision making process. The disadvantage of these indices is that they do not take into account the complexities of conserving multiple taxa. For example, if one taxon is conserved the relative importance of conserving closely related taxa may decrease as we wish to conserve as distinctive a *set* of taxa as possible.

Prioritising species using species specific indices differs substantially from using PD maximising approaches. PD is uninformative for any one species on an ultrametric tree – all single species are the same distance from the root and so receive the same value. Many current conservation approaches (e.g. endangered species lists) rely on having species ranked in order of priority. Current PD approaches offer no such order. To overcome this, there is a natural index value associated with a given PD maximising solution, which we introduce here.

Consider adding all n species in a tree to a set, S, one at a time using the greedy algorithm in Theorem 7. A natural index value to associate with a species, i, is the additional PD that i contributes to S when it is added to S. We denote the sequence of sets by $S_1, S_2, \ldots, S_{n-1}, S_n$ with $|S_j| = j$ and $S_j \subset S_l,\ \forall\ j < l \leq n$. The index value associated with a given species, $\urcorner(i)$, is the additional PD that species i brings to the set when it is added:

$$\urcorner(S_{j+1} - S_j) = PD(S_{j+1}) - PD(S_j).$$

The problem with $\urcorner(i)$ is that for a given problem the sequence of sets

$S_1, \ldots, S_n$ may not be unique (especially if the tree is ultrametric) and the index values may therefore not be unique either. In fact the number of possible index values for each species may actually exceed the number of PD maximizing solutions. This will again make it difficult to utilise this approach at the management level.

More importantly, the amount of PD saved is only optimal if all the species that are selected are subsequently protected. If any species in the selection are lost, the remaining species may be far from optimal. Finally, it may be difficult to find optimal sets of species if there are large numbers of species to prioritize, and other complex factors such as cost of conserving individual species are considered (see chapter 7).

The species specific measures of evolutionary distinctiveness we present are a flexible and transparent tool for including 'evolutionary value' in conservation management. However, and importantly, they have not been designed to capture total PD. If sets of evolutionary distinctive species did capture substantial PD, then the species-specific measures would be doubly useful, highlighting the most individually distinctive species and helping preserve more of the tree of life. In this chapter we compare the PD of sets of species selected using the species specific indices with the PD of sets of species obtained using the PD maximising algorithm described in chapter 6. We also consider aspects of the underlying tree that affect the amount of PD captured using the species specific indices.

10.1 The indices

Here we consider some simple indices with a particular focus on those that utilise a phylogenetic tree. For the interested reader some notable work not considered here is contained in Barker (2002), Clarke and Warwick (1998), Crozier (1992), Haake et al. (2008) and Vane-Wright et al. (1991).

10.1.1 Pendant edge

One of the conceptually simplest indices is the 'Pendant Edge' (PE) measure introduced in Altschul and Lipman (1990) where each taxon is assigned a value equal to the length of its pendant edge. The PE value of each species

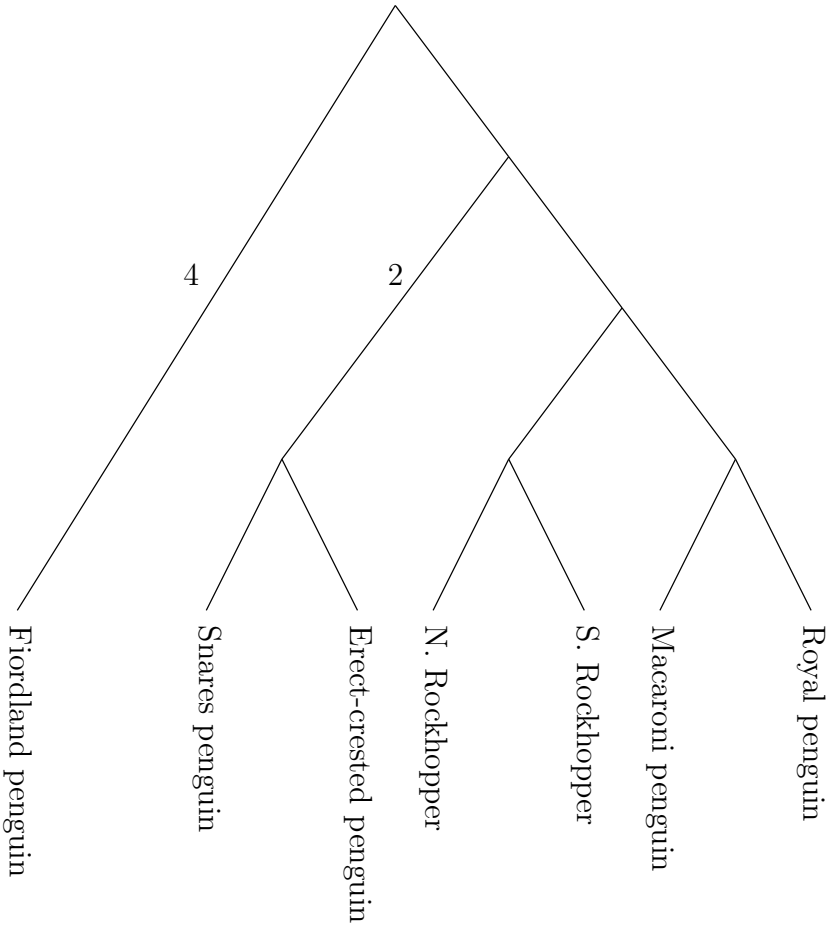


Figure 10.1: The phylogenetic tree for crested penguins. This tree was derived from the tree in Bertelli and Giannini (2005) and Giannini and Bertelli (2004) which had no branch lengths. For illustrative purposes each level in the original tree was assumed to be separated by the same distance such that all edges in this tree are of length 1 except for the two marked edges.

in Figure 10.1, is easily determined: in this case the Fiordland penguin has the highest PE with the other taxa having an equal second highest value. PE suggests that the Fiordland penguin is the most important to conserve but does not differentiate between the other taxa. It seems logical, though, that if some of the other taxa were to be conserved, we should not choose the most closely related of these.

10.1.2 Shapley value

One can also consider the expected contribution to PD that a taxon will make at some time in the future if the survival of all other taxa is uncertain. To make this idea precise, for each subset S of X, and each taxon $i \in X - S$ let

$$\Delta_{PD}(S, i) = PD(S \cup \{i\}) - PD(S),$$

where $\Delta_{PD}(S, i)$ is the increase in PD that taxon i provides when added to S. Now, suppose that each taxon $j \in X - \{i\}$ has a probability a_j that it is not extinct at some time t in the future. If we assume that extinction events are independent between taxa, and let E be the (random) set of taxa that are extant at time t, then we can ask how much we expect taxon i to contribute to the PD at time t. This contribution, ψ_i is simply the expected value of $\Delta_{PD}(E, i)$, given formally by

$$\psi_i = \sum_{S \subseteq X - \{i\}} \mathbb{P}[E = S] \Delta_{PD}(S, i).$$

Note that $\mathbb{P}[E = S]$ is the probability that the set of extant taxa at time t will be S.

The special case of ψ_i where $\mathbb{P}[E = S] = \frac{(|S|-1)!(n-|S|)!}{n!}$ is the Shapley value which originated in game theory and has been considered in detail in a phylogenetic context by Haake et al. (2008). In this case all sets of a given size are equally likely and a number of 'fairness' axioms are satisfied (Haake et al., 2008). An ecologically based variant where $\mathbb{P}[E = S]$ depends on species' survival probabilities is presented in Steel et al. (2007).

10.1.3 Equal splits and fair proportion

A conceptually appealing family of indices divides the total phylogenetic diversity of a tree amongst the taxa corresponding to the leaves of that tree. An example of these indices is the Equal Splits (ES) index (Redding and Mooers, 2006) which is closely related to the previously discussed Pauplin formula (Equation 6.1). This index splits the length of an edge equally between its daughter trees. Denoting the edge length between a node, j, and its direct ancestor by λ_j, the equal splits index for a taxon, i, can be calculated by summation over all the nodes between i and the root (including i):

$$ES(i) = \sum_j \frac{\lambda_j}{2^{d'(i,j)}} \tag{10.1}$$

where $d'(i,j)$ is the number of edges between the taxon (node i) and node j. Applying the ES index to the tree in Figure 10.1 again suggests that the Fiordland penguins are the most important species to conserve with an index value of 4. The Snares and Erect-crested penguins have an index equal value of $2\frac{1}{4}$ whilst the remaining species have a value of $1\frac{7}{8}$; if for example three species could be conserved this suggests that the Fiordland, Snares and Erect-crested penguins should be chosen. Intuitively, however, it seems more beneficial to conserve one of the other species instead of the Snares or Erect-crested penguins and thus protect more of the internal edges. The problem with ES (and other simple indices) is that the decision to conserve one taxon does not affect the importance assigned to conserving the remaining taxa.

The Fair Proportion (FP) index is a variation of the Equal-splits index where each taxon descendant from an edge is allocated an equal proportion of that edge length. Formally:

$$FP(i) = \sum_j \frac{\lambda_j}{C_j}.$$

This differs from the ES index purely in the proportion of an edge that is allocated to each of its descendants. The advantage of FP is that, as we will show in this chapter, it is closely related to the Shapley value.

10.1.4 Quadratic entropy

A further index that has seen some application is Quadratic Entropy (QE; Rao (1982); Pavoine et al. (2005)). The QE index is calculated by maximising:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(QE(i)QE(j)\sum_{k\in P(i,j)}\lambda_k\right),$$

subject to the constraint that the QE indices are non-negative and sum to unity. Here we have denoted the set of edges connecting species i and j by $P(i,j)$. In Bokal an efficient algorithm for computing the QE indices has been produced. The conceptual complexity of QE reduces its advantage over PD approaches as it would be difficult to explain to conservation managers and stakeholders. In this chapter we also show that it has the least utility for conservation management out of the considered indices, hence we limit further discussion of QE.

10.2 Species specific indices versus PD

Consider the situation where we wish to find an optimal set of k species for a given tree. This can be achieved with a species specific index by calculating the index value and selecting the k species with the highest values. Indices that produce sets of species with high PD in this manner are likely to be of greatest use for selecting biodiverse sets of species when incorporated into the more complex conservation management frameworks already in use. Redding et al. (2008) considered the PD represented by sets constructed in this way using some of the indices introduced in the previous section.

Here we present a very similar analysis to that in Redding et al. (2008), placing greater emphasis on the relationship between the species specific indices and PD. The 'real' trees considered are the 46 near-ultrametric trees from McPeek and Brown (2007) discussed in chapter 9 and 40-species Yule trees that were produced using the methods described in chapter 3.

Figure 10.2A shows the proportion of the total PD in the 'real' trees represented by sets of species selected using the different indices to prioritise

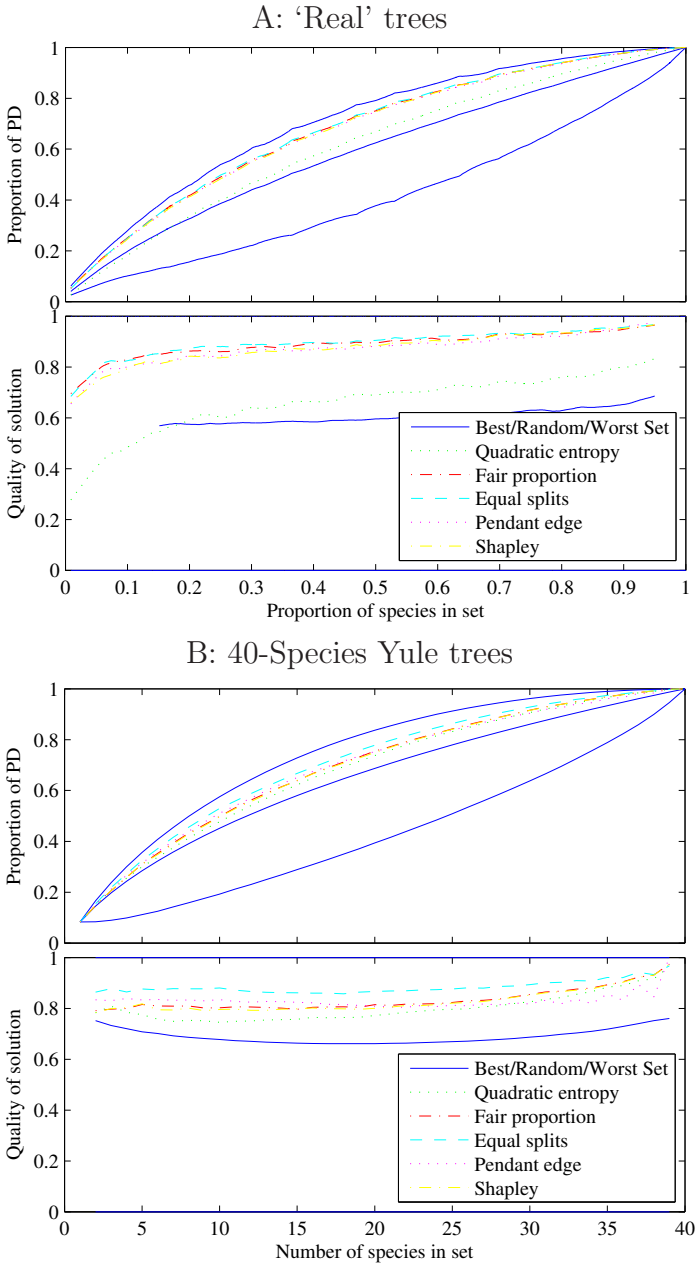


Figure 10.2: The proportion of PD represented by sets of species for both (A) the 'real' trees and (B) 40-species Yule trees. The sets were selected by prioritising species according to the various indices. Also depicted are the best, random and worst sets. The bottom graph in each panel shows the quality of the solution where solutions of quality zero and one respectively have the lowest and highest possible PD. The blue line in the bottom panels, therefore corresponds to the random sets (as the best and worst sets always have values one and zero and are therefore not displayed). The curves in (A) are Loess smoothed as there were many different sized trees and the x-values therefore vary between trees.

them. The PD represented by random sets of species and the PD maximising and minimising sets was also calculated using the methods described in chapter 6. The lower graph in Figure 10.2A shows the quality of the solutions for the same sets; this is a rescaling such that a set with quality one is a PD maximising set and a set with quality zero is a PD minimising set. All the indices outperformed random species choice, ES produced the best solutions although apart from QE the other indices were close to ES.

Figure 10.2B shows the same analysis for 40-species Yule trees. In this situation ES produces notably better solutions than the other indices and QE does not perform as poorly.

Overall these analyses suggest that ES is the best index for identifying sets of species with high PD. The other indices performed well except for QE which should be avoided due to its poor performance for the 'real' tree dataset.

In the following sections we provide a mathematical explanation of (i) why random choice captures a large amount of PD (ii) why ES and PE perform better than random choice (iii) why there is an extremely strong correlation between FP and the Shapley value.

10.2.1 Random choice

Random choice captures a large proportion of total PD (Nee and May, 1997), however in Redding et al. (2008) it was shown that random choice performs poorly with increasing tree imbalance. To the extent that published trees are more imbalanced than Yule trees (Mooers and Heard (1997), Blum and Francois (2006) chapter 4), random choice is compromised.

Consider an ultrametric tree representing n species of which some number, k, can be conserved. All the possible sets of k species with maximal PD can be represented graphically using the following simple method. Locate a time, l, at which k ancestral species exist and partition the n species into the k subsets descendant from these ancestral species. We will refer to these as the optimal k clades: all optimal solutions correspond to selecting one species from each of these clades (Nee and May, 1997).

We first ask: How many of the k clades will be represented by a random

selection of species (with the optimal solutions corresponding to all of these clades being represented)? A clade is considered to be represented if at least one species from that clade is conserved. We denote the probability that a clade of size i is represented by ϕ_i, which corresponds to one minus the probability that all of the k randomly selected species will be chosen from other clades:

$$\phi_i = 1 - \frac{\binom{n-i}{k}}{\binom{n}{k}}.$$

where $\binom{n}{k}$ is the number of ways of selecting k elements from a set of n elements. For brevity, we adopt the usual convention that $\binom{n}{k} = 0$ if $k > n$. We denote the number of the k clades containing i species by a_i. Using this notation the expected number of the k clades that will be represented by a random selection of species is easily expressed:

$$\hat{N} = k - \sum_i \frac{a_i \binom{n-i}{k}}{\binom{n}{k}}. \tag{10.2}$$

The expected number of clades that are represented depends on the size distribution of the clades (a). The size distribution for which the least number of clades are expected to be represented is that where all clades contain a single species except for one that contains $n-k+1$. For this size distribution, Equation 10.2 simplifies to:

$$\begin{aligned} \hat{N} &= k - \frac{(k-1)\binom{n-1}{k} + \binom{k-1}{k}}{\binom{n}{k}} \\ &= 1 + \frac{k(k-1)}{n}. \end{aligned} \tag{10.3}$$

The highest proportion of the k clades is expected to be represented if the species are as evenly distributed amongst the k clades as possible, i.e. if the tree is completely balanced. If n is a multiple of k, Equation 10.2 simplifies to:

$$\hat{N} = k - \frac{k\binom{n-n/k}{k}}{\binom{n}{k}}. \quad (10.4)$$

Equations 10.2 and 10.4 provide a lower and upper bound on the expected proportion of the clades that will be represented by randomly selecting species for any tree. Under the Yule process, the distribution of the sizes of the k clades is geometric (Nee et al., 1992), such that we are nearer the lower than the upper limit (Figure 10.3). Note that Hey trees use the same branching patterns as Yule, so any arguments based on tree topology should be applicable to both tree types.

10.2.2 Pendant edge

Using the notation introduced above, when selecting species using PE, all of the k clades with size one are automatically represented as they have longer pendant edges than species from any multi-species clades. Hence those clades that are the least likely to be represented by random species selection are guaranteed to be represented under the PE measure. This means that PE will do better than random on the most unbalanced tree shapes. However, we must still ask how well the k clades of size ¿ 1 are represented under the PE measure. If more than one species is picked from one of these larger clades, some of the other clades must be unrepresented. It is clear that PE will represent more clades if the difference between the longest pendant edges and the other edges in each clade is as big as possible.

To examine this we considered several aspects of edge length probability distributions. These distributions can be obtained by simulating Yule tree data (as described in chapter 3); however, this is a time consuming approach, as large numbers of trees must be simulated to obtain reliable results. Here, we use the analytic methods described in chapter 2.

This allows us to make the following observations on Yule trees: first, though the actual topology of the subsets in k affect the distribution of PE among k, larger subsets from k are expected to have a longer PE than smaller subsets from k, and therefore, are more likely to be represented (Figure 10.4).

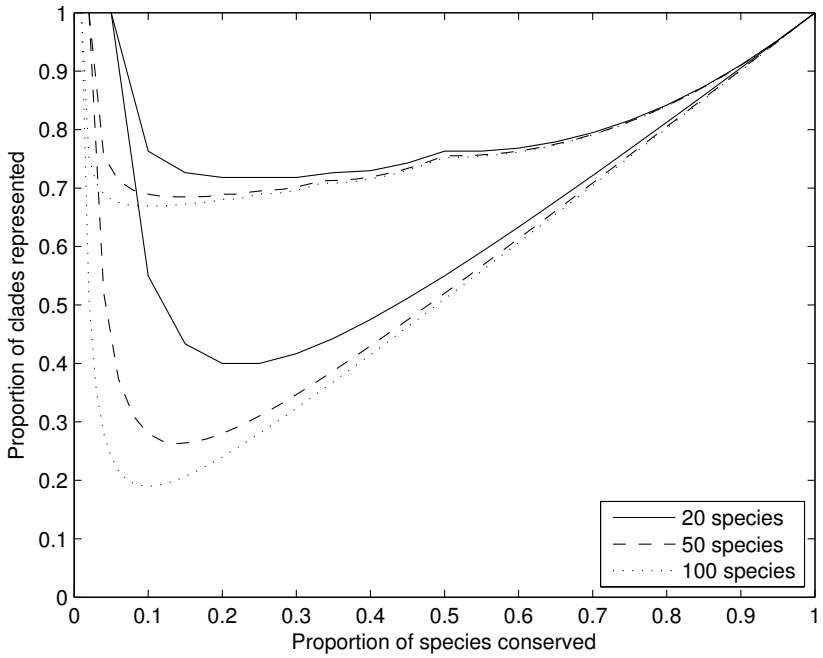


Figure 10.3: Expected proportion of the optimal PD captured by random species choice on Yule trees. Under optimal choice, each of k clades (where k is the number of species conserved) should be represented (see text for details). So, for $n = 20$ species, if 5% of species chosen (= 1 species), this must also capture the maximum number of clades, since $k = 1$. Likewise, if all species are chosen, then every one of $k = n$ optimal clades must also be represented. At intermediate values, random species choice will represent $< k$ clades, and this deficit increases with tree size.

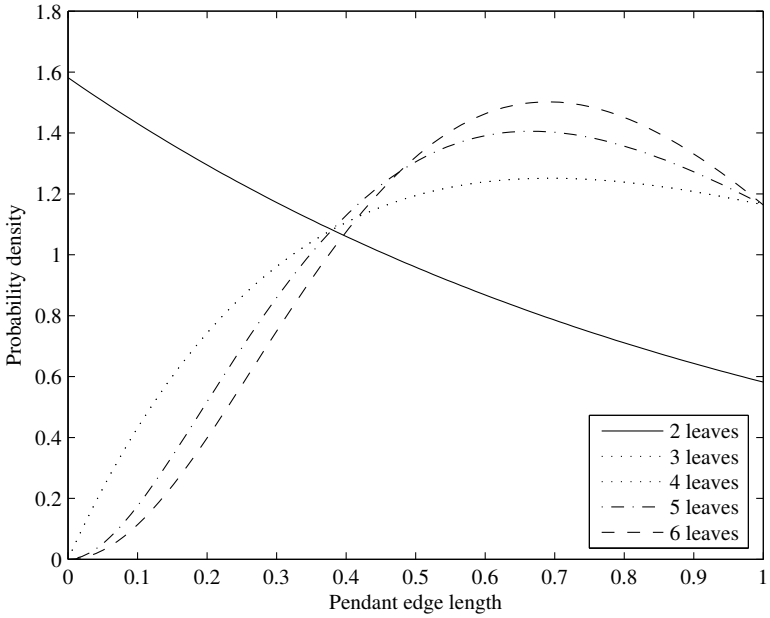


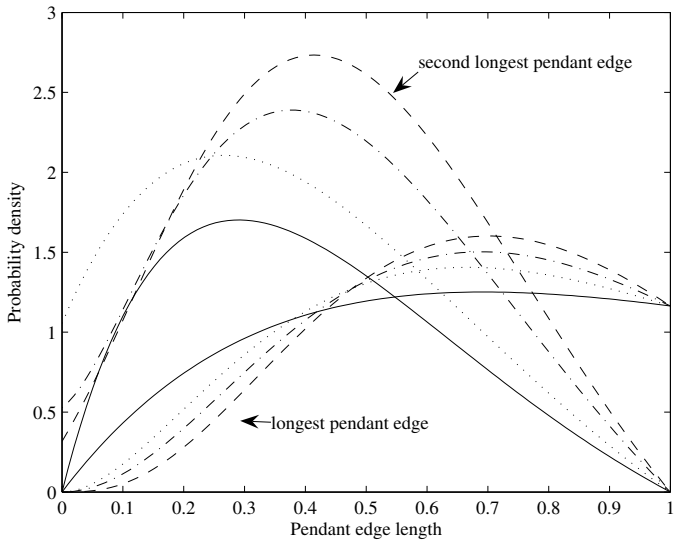
Figure 10.4: The probability density of the longest pendant edge for a range of subtree sizes: the density shifts right for larger sub-tree sizes, illustrating that the maximum pendant edge length is more likely in larger sub-clades. Note that sub-trees with three and four species have exactly the same density.

In this way, PE will act like random choice. In addition, some of the second-longest edges in some clades may be longer than the longest PEs in other clades, which will lead PE to capture less PD (Figure 10.5A). That said, there is good contrast between the longest PEs and the remaining PEs (Figure 10.5B), implying that, in general, PE will not choose repeatedly from the same clade.

10.2.3 Equal splits

The ES measure incorporates the PE measure. In fact, the pendant edge generally contributes over half the ES score on Yule trees. On a fully bifurcating tree, if we denote the edges between a pendant edge and the root by λ_0 through to λ_r, the ES measure of a species (Equation 10.1) can be reduced to:

A



B

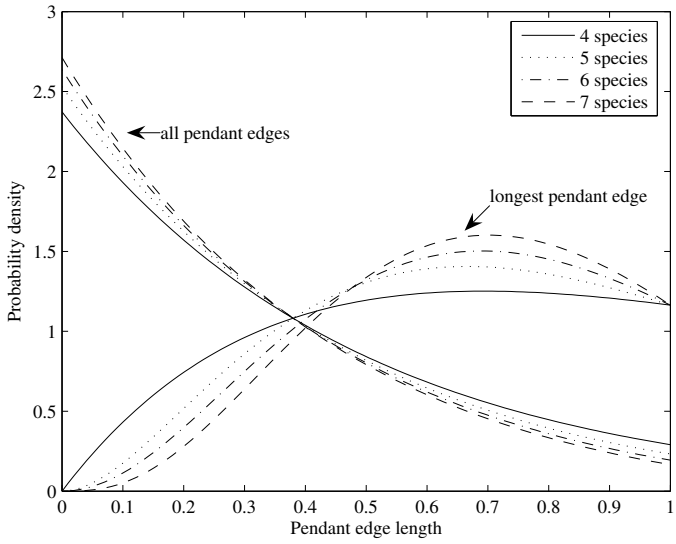


Figure 10.5: The ability of pendant edge (PE) measure to capture PD on Yule trees. PE will fail to the extent that it chooses repeatedly from one of the k optimal clades (see text for details). (A) Probability density of the longest (solid lines) and second longest pendant edges (dashed lines) in one of the k clades. The overlap of the two distributions illustrates that it is likely for the sub-tree to be sampled twice by PE. (B) The solid curve is again the probability density for the longest pendant edge, while the dashed line is for all the other pendant edges in a sub-tree. The contrast between the two curves highlights why PE generally over-samples single sub-clades less than random choice.

$$ES = \sum_{i}^{r} \frac{\lambda_i}{2^i}.$$

If the branch lengths were equal this is simply a geometric series and we obtain:

$$ES = 2\lambda \left(1 - \frac{1}{2^{r+1}})\right) \leq 2\lambda.$$

The extent to which pendant edges affect the ES measure depends on the relative magnitudes of pendant edges and interior edges. If these are of similar magnitude, the pendant edges will on average contribute half of the ES value. For pure Yule trees, PEs are on average slightly shorter than internal edges as they represent the time from the birth of a species to the present, not the time from the birth of a species to a speciation event.

In any of the k clades, the expected number of internal nodes between a pendant edge and the rest of the tree is lowest for the longest pendant edge due to topological constraints (consider a ladder, or fully imbalanced tree). This increases the contrast between the maximal ES score in each of the k clades and the other ES scores in those clades, when compared to the contrasts found between PE scores in the same situation. This, therefore, directs ES species choice more efficiently to smaller clades and reduces the number of clades in the tree that are over-represented.

In summary, the ES measure is highly related to the PE measure and for Yule trees is expected to represent a larger proportion of the k clades than the PE measure alone. However, the biases that affect the PE measure are expected to also affect ES, due to their correlation. Findings applicable to ES are also likely to be similarly applicable to FP, due to their mathematical similarity. It is unknown why ES appears to outperform FP; further study is needed to investigate this property.

10.3 Relationship between Shapley and Fair Proportion measures

From the analysis presented here (Figure 10.2) and in Redding et al. (2008) it is apparent that there is a high degree of correlation between the Shapley

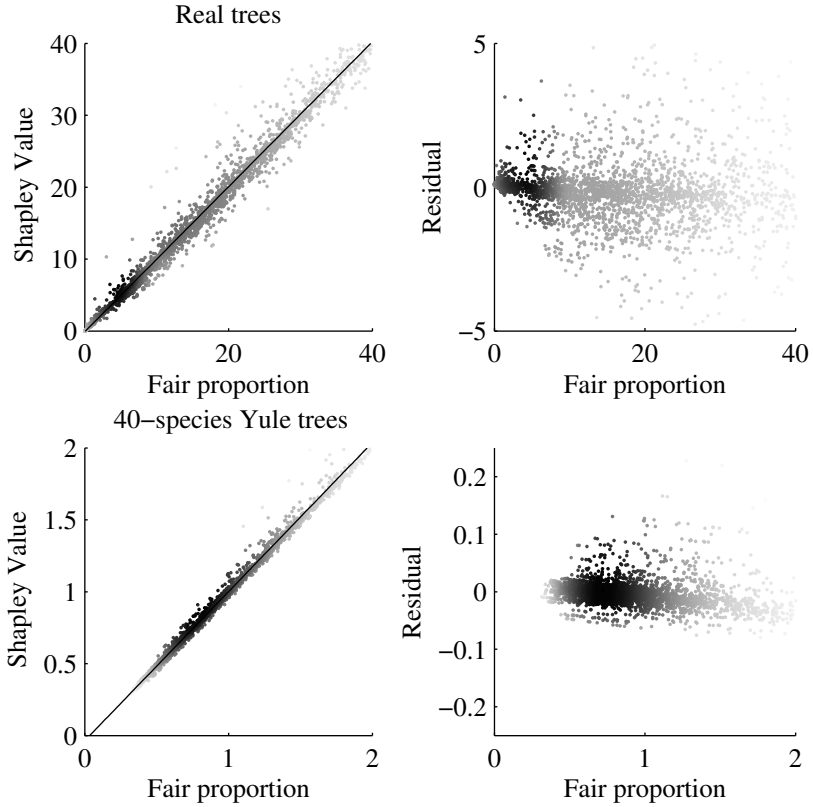


Figure 10.6: The panels on the left show a scatter plot of Shapley values against the corresponding fair proportion measure for each species in the sample of trees. Darker colours correspond to a higher density of points. The lines depict a linear regression of the two values (see Table 10.1). The panels on the right show the residuals resulting from this regression.

and Fair Proportion indices. In Figure 10.6 we show the relationship between these indices for the previously considered 'real' and Yule trees. A linear regression between these two measures was conducted and an extremely strong relationship was observed (see Table 10.1). This is of particular interest as the fair proportion measure has been (somewhat arbitrarily) selected for the EDGE of Existence conservation program (Zoological Society of London, 2008). The Shapley value has a nice practical interpretation which due to this result can also be attributed to the fair proportion measure.

In the remainder of this section we will show the mathematical underpinning for the close relationship between these indices. Consider the contribu-

Tree sample	Constant	Gradient	r^2
'Real' trees	-0.10 ± 0.05	1.008 ± 0.002	0.99
40-species Yule trees	-0.035 ± 0.003	1.037 ± 0.003	0.99

Table 10.1: Coefficient estimates (with 95% CI) and r^2 values for a linear regression of the Shapley value against the fair proportion measure.

tion an edge, e makes to the index value of a given species, a. For the fair proportion index this is simply:

$$\theta_a(e) = \begin{cases} \lambda_e / C_e & a \in \mathcal{C}_e \\ 0 & a \notin \mathcal{C}_e, \end{cases}$$

where $\mathcal{C}_e$ denotes the set of children of edge e and $C_e = |\mathcal{C}_e|$.

For the Shapley value this becomes more complicated. Let J denote the set of species separated from a by the edge e (see Figure 10.7). Recall that the Shapley value is expressed as:

$$\phi(a) = \frac{1}{n!} \sum_{S, a \in S} (|S| - 1)!(n - |S|)! \left(PD(S) - PD(S - \{a\}) \right).$$

An edge contributes to the Shapley value of species a if all species in S are separated from a by the edge e: $S - J = \{a\}$. Note that an edge can only contribute to $\phi(a)$ if $2 \leq |S| \leq |J| + 1$. For a given subset size, $|S|$, there are $\binom{|J|}{|S|-1}$ sets where this occurs. The total contribution that e makes to species a is its coefficient in Equation 10.3 times the number of sets to which a makes a contribution:

$$\begin{aligned} \psi_a(e) &= \sum_{|S|=2}^{|J|+1} \lambda_e \frac{(|S| - 1)!(n - |S|)!}{n!} \times \binom{|J|}{|S| - 1} \\ &= \sum_{|S|=2}^{|J|+1} \lambda_e \frac{|J|!(n - |S|)!}{(|J| - |S| + 1)!n!}. \end{aligned} \tag{10.5}$$

Theorem 16. *For a constant C_e in the limit as $n \to \infty$ we have $\psi_a(e) = \theta_a(e)$. In other words the contribution of edge e to the Shapley value and FP*

value of species a becomes equivalent.

Proof. The contribution that an edge e makes to the Shapley value of a species in J is found by substituting $n - |J|$ for J in Equation 10.5. The contribution that e makes to all the species in J is found by making this substition in Equation 10.5 and summing the result over all species in J:

$$\begin{aligned}\sum_{i \in J} \phi_a(e) &= \lambda_e |J| \sum_{s=2}^{n-|J|+1} \frac{(n-|J|)!(n-s)!}{(n-|J|-s+1)!n!} \\ &= \lambda_e \sum_{s=2}^{C_e+1} \frac{|J| C_e!(n-s)!}{(C_e-s+1)!n!}\end{aligned}$$

taking the limit we obtain:

$$\begin{aligned}\lim_{n\to\infty} \sum_{i \in J} \phi_a(e) &= \lambda_e \lim_{n\to\infty} \sum_{s=2}^{C_e+1} \frac{C_e!(n-s)!|J|}{(C_e-s+1)!n!} \\ &= 0.\end{aligned}$$

For the Shapley value it is known that each edge is shared out in its entirety amongst the species (Haake et al., 2008), therefore λ_e will be divided amongst the species in $\mathcal{C}_e$. From Equation 10.5 each species in $\mathcal{C}_e$ has the same value of $\psi_a(e)$ which must therefore be λ_e / C_e as required.

The extent of the similarity between the Shapley value and FP for a species depends on the number of edges separating the species from the root and the number of species contained elsewhere in the tree.

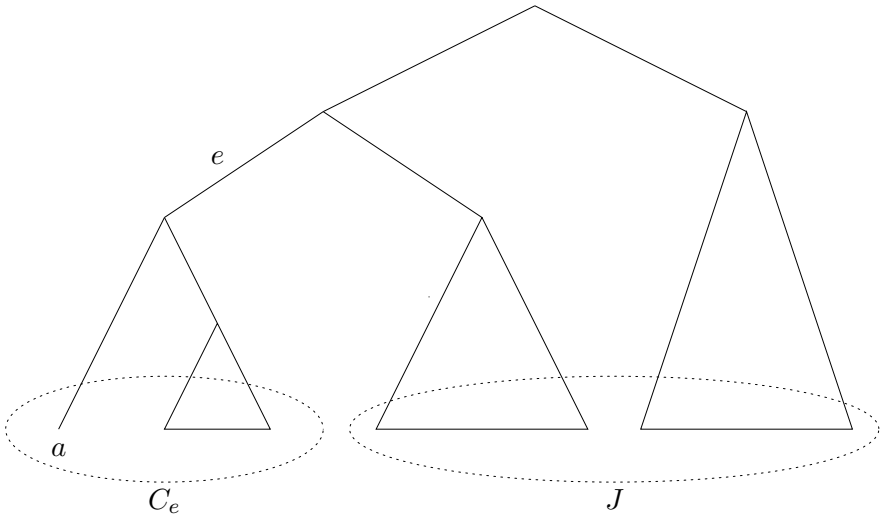


Figure 10.7: Edge e separates the species in J from the set of species below it – $\mathcal{C}_e$.

Chapter XI

Future directions

The magnitude of the current extinction event is enormous. Combined with the growing pressure humans are placing on our ecosystems, conservation is more important than ever before. Resources available for conservation are limited, therefore we should ensure they are allocated as efficiently as possible.

As shown in chapter 9, including phylogenies in conservation management can result in substantial biodiversity gains. In the previous chapters we have provided new results for several methods that can be used to achieve this. In addition to the work in this thesis, there is a substantial body of theoretical literature concerning these approaches, however their application has remained limited. To overcome this, further effort needs to be made to apply these scenarios in practical situations and to address some of the further practical concerns of these methods.

In this chapter we briefly outline some promising future directions of research in this field that may help bridge the gap between the theoretical literature and its application. The topics discussed here are both important for practical conservation applications and present a logical continuation of the work presented in this thesis.

11.1 Is PD appropriate?

Throughout this thesis it has been assumed that PD is an appropriate measure of biodiversity; indeed non-PD methods have been evaluated by the amount of PD they capture. In this section we consider whether a phylogenetic tree is an appropriate structure on which to measure biodiversity and if so whether PD is the appropriate measure to use on this structure.

As previously noted the edge lengths in a phylogenetic tree can have many different interpretations (eg. genetic distance, time, function distance) depending on the data from which the tree was constructed. Most applications of PD to date have utilised trees that have been constructed from genetic sequences or morphological data. These trees have been constructed to provide an approximation of the evolutionary pathways that have given rise to the modern species. For such trees PD maximising sets will have high genetic diversity. For the purpose of conservation, however, the functional traits and ecosystem role of species may be more important than genetic diversity.

Petchey and Gaston (2002) considered trees constructed from functional traits which they refer to as functional dendograms. They applied PD to the functional dendograms and called the resulting measure Functional Diversity (FD). We find this a confusing label, as it is the data from which the tree has been constructed that has changed (functional traits instead of eg. genetic sequences) and not the measure on the tree; hence we refer to this measure here as PD on functional dendograms. In terms of ecosystem function, it may be more appropriate to consider PD on functional dendograms than on evolutionary trees.

An important question is to what extent functional traits are inherited and consequently how similar a functional dendogram and the corresponding evolutionary tree are. It would be an interesting exercise comparing species prioritisations for situations where both are available. Indeed, a tree structure may even be inappropriate for functional traits, if the functional dendogram does not exhibit a high degree of topological similarity with the evolutionary tree.

The next question is whether PD is an appropriate measure, regardless of the data from which the tree is constructed. PD prioritises evolutionarily distinct species, however the most distinct species could be evolutionary dead ends – commonly referred to as living fossils. Furthermore a clade containing many closely related species may be an indication that the functional role of those species is of high importance or that it is a promising direction for evolution. Consider Figure 11.1, under PD the species from the circled species rich clade have equal or lesser importance than those in the other clades, however for ecosystem function these species may in fact be the most

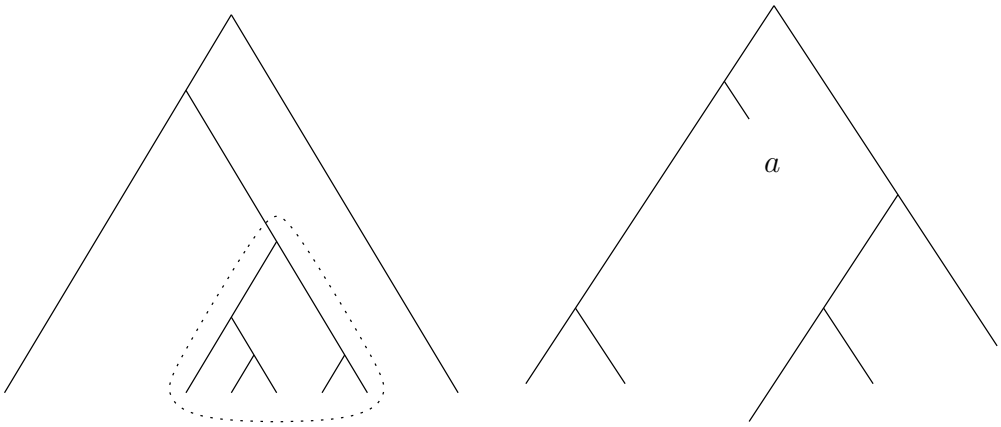


Figure 11.1: Left: The PD measure will give the same importance to the species in the circled species rich clade as to the others. However this rich clade may provide a more important functional role and should therefore be assigned greater importance. Right: Species a diverged a long time ago and has evolved little since that time. PD would assign little importance to a. The species in the cherry that is sister to a would be selected first and a then provides only a minimal increase in PD. From a scientific point of view a may provide a unique window to the past and should perhaps be conserved for that reason.

important. If we could only conserve two species perhaps we should conserve two species from the large clade, a direct contradiction to PD.

On non-ultrametric trees, PD gives minimal importance to basal species that have undergone minimal functional evolution since speciating (Figure 11.1). Such species (e.g. Tuatara) provide a unique glimpse into the past and may be of great scientific value.

The last two paragraphs may seem somewhat contradictory, first we state that greater emphasis should be placed on closely related species and then, that it should be placed on evolutionarily distinct species. The reason for this apparent contradiction is that two different justifications for selecting species are used. Which (if either) of these justifications are better is a study beyond the scope of this thesis and will depend on the application. A better alternative may be to retain PD as a measure of biodiversity and include other factors (eg. the functional role of a species) separately. One such approach that seeks to maximise PD whilst retaining a viable food web is

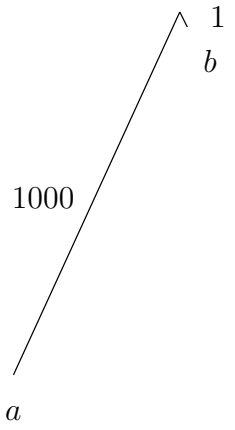


Figure 11.2: Here species a has a pendant edge that is a thousand times that of species b (edges are not to scale). Consider the situation where conserving species a raises its survival probability from 0 to 1% and conserving species b raises its survival probability from 0 to 100%. The optimal NAP solution with a single species would include a despite the fact that with 99% probability this will result in zero PD. A more risk averse management strategy would select species b.

outlined in Moulton et al. (2007).

11.2 Are 'optimal' solutions best?

Throughout this thesis we have sought to produce solutions that maximise quantities such as PD. The problem with this is that unlikely outcomes with low biodiversity can still occur. An extreme example would be a situation where a species, a, is a thousand times more valuable than species b. Furthermore species a has a 1% chance of being saved if the entire conservation budget is spent on it, whereas species b could be saved with certainty (see Figure 11.2). The expected value of conserving species a is ten times that of conserving species b although the most likely outcome in this scenario is that neither species is saved. In many real conservation problems the desirable course of action may in fact be to conserve b.

An alternative to maximising the expected PD would be to maximise the worst possible outcome. However unless some species are guaranteed to survive, a zero PD solution is always possible. A more appropriate option would be to ensure that, in say 95% of cases, the PD will exceed a minimum threshold that is as high as possible. We let $f(PD|S)$ denote the probability density of the PD obtained by conserving the species in S and $F(PD|S)$ the cumulative density. Previously we have maximised $\mathbb{E}(f(PD|S))$ (although we referred to it as $\mathbb{E}(PD|S)$), a suitable alternative for risk averse management

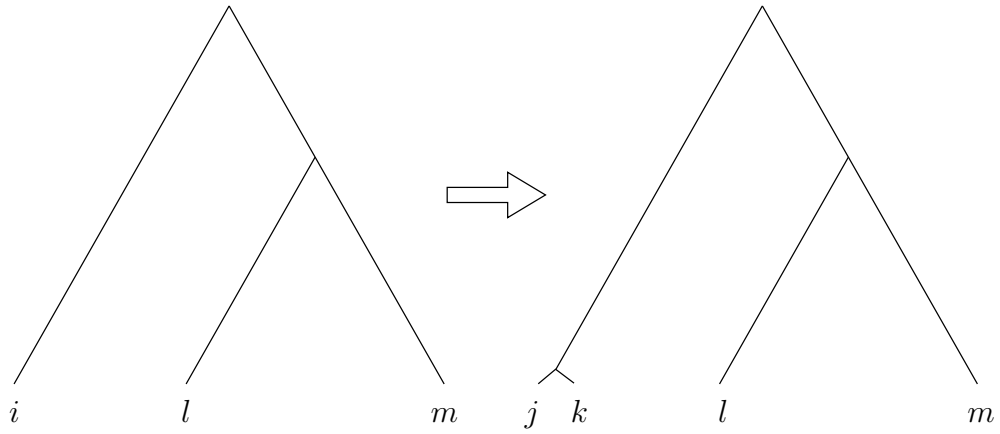


Figure 11.3: The initial accepted phylogeny for a tree is shown in the left panel. New data then reveals that species i is in fact two very closely related species – j and k. Ideally a prioritisation method such as the NAP should be robust to such changes.

is to maximise:

$$\text{maximise}_S F^{-1}(\gamma|S).$$

If γ is high, management is optimistically maximising the best possible outcomes and in our previous example would select species a. If γ is low, management will be risk averse and would select species b. Note that setting $\gamma = 0.5$ will maximise the median outcome. Choosing an appropriate γ value is a complex problem, it has no clear answer and is beyond the scope of this thesis.

Developing efficient methods for this maximisation problem is likely to be challenging, especially in the context of the NAP. For some initial work on characterising the distribution of PD for given survival probabilities see Faller et al. (2008).

11.3 What happens when a species is split in two?

As discussed in chapter 10, the measures used throughout this thesis are highly dependent on the length of pendant edges. Unfortunately pendant edge lengths are susceptible to dramatic change through the discovery of new species – particularly when a single species is resolved into two closely related

species. Consider the situation in Figure 11.3. Initially there is a single species, i, with NAP parameters a_i, b_i, c_i. This species is then determined to actually be two very closely species, j and k (Figure 11.3). Since the new species are very closely related one might desire that the combined budget allocation to j and k (from the optimal NAP solution) would be equal to the original allocation to i. We now consider one possible set of parameter restrictions that ensure that this is achieved. Firstly, the probability of both j and k becoming extinct should be equal to that of i (such that the scaling in chapter 7 remains the same):

$$\begin{aligned} 1 - a_i &= (1 - a_j)(1 - a_k) \\ 1 - b_i &= (1 - b_j)(1 - b_k) \end{aligned}$$

Secondly, the cost of conserving both new species should be equal to that of the original species:

$$c_j + c_k = c_i.$$

For parameters that satisfy these restrictions *and* for which the parameters for j and k have equal values, the solutions to the NAP will indeed be equivalent. However if j and k have different parameter values the NAP solutions may differ. This is because one species will provide a greater biodiversity gain per unit of resource than the other. The importance of the two species will consequently be altered.

For the NAP this situation can be resolved in the manner we have just described, however for species specific indices no such simple solution is available. Recall from chapter 10 that the pendant edge of a species contributes roughly half the index value for the considered indices. For the indices considered in that chapter, splitting a species will roughly halve the index value for the new species (eg. $ES(i) = ES(j)/2 = ES(k)/2$). For the pendant edge measure this is even more pronounced as the measure is set to zero. The new species (j and k) will therefore have equal index values which are significantly than that of the original species (i). This will significantly alter the species prioritisation, for example a single important species may be transformed into two unimportant species.

The sensitivity of the indices to the splitting of species is a major drawback and it is unclear how to correct this without affecting the simplicity of the indices (their major advantage over PD).

In reality the assumption that the two new species would be separated by a zero length edge will clearly not be satisfied. In general we would expect that the two species will split the pendant edge of the original species, resulting in higher PD in that portion of the tree. Lastly we note that discovering that a single species is actually two, may completely change the practical conservation approach for those two species. The actual combined conservation costs and survival probabilities for the two new species may be drastically different to that for the original single species.

11.4 How should artificial polytomies be handled?

Consider a tree that contains polytomies that are due to lack of information and not a true multiple speciation. Depending on the method by which the tree was constructed these polytomies generally increase the total PD contained in the tree. This is particularly problematic as the biodiversity in the tree will become biased towards clades containing polytomies.

Here we consider one method for rescaling branch lengths in a tree that corrects the total PD. Under this correction indices that assign the total PD of the tree among its leaves will obtain their expected value given our uncertain knowledge of the true tree. This is due to the linearity of the indices considered in this thesis.

Consider a single polytomy as illustrated in Figure 11.4. This polytomy has j descendant trees $(\mathcal{T}_1,\ldots,\mathcal{T}_j)$, hence we call it a polytomy of degree j. There is some time, t, during which the $j-1$ speciation events producing the j trees may have occurred. If the time of the polytomy denotes the first time at which one of the descendant trees may have split off then t is simply the time between the polytomy and the first speciation event in one of the daughter trees.

Given our definition of t we can look at the component of the tree of length t following the polytomy. This component at present has a PD of jt since j lineages are present for t time units. In actuality the j lineages

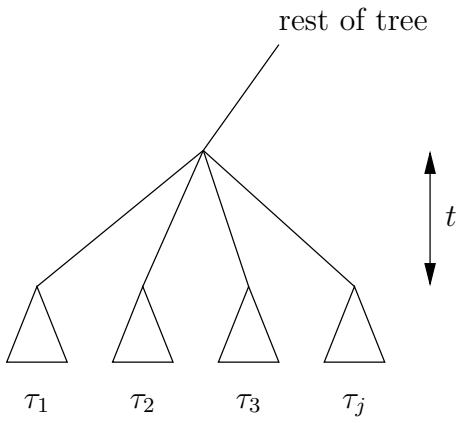


Figure 11.4: A single polytomy occurring somewhere in a tree. The polytomy has j descendant species. We assume that the $j-1$ speciation events occurred sometime during the interval t. Alternatively we could assume that the first speciation event occurred exactly at the time of the polytomy followed by $j-2$ speciation events during the interval t.

were not all present for this length of time, hence the PD of this component should be lower.

Here we provide a method for using evolutionary models to calculate the expected PD of the component. For simplicity we assume that the evolutionary model is memoryless and independent of the total number of species present at that time. Hence the species ancestral to the polytomy can be considered the start of a new tree which at age t has a speciation event resulting in the $(j+1)$th species. The corrected PD of the component is therefore the expected PD of a tree under this evolutionary model with the $j+1$th speciation event occurring at time t.

The total PD of a tree is determined only by the duration for which different numbers of species were present, not by the tree shape. Denoting the time for which there were $s_2, \ldots, s_j$ species the PD of a tree with j species is:

$$PD = \sum_{i=2}^{j} i s_i$$

The expected PD of the component under a given evolutionary model is

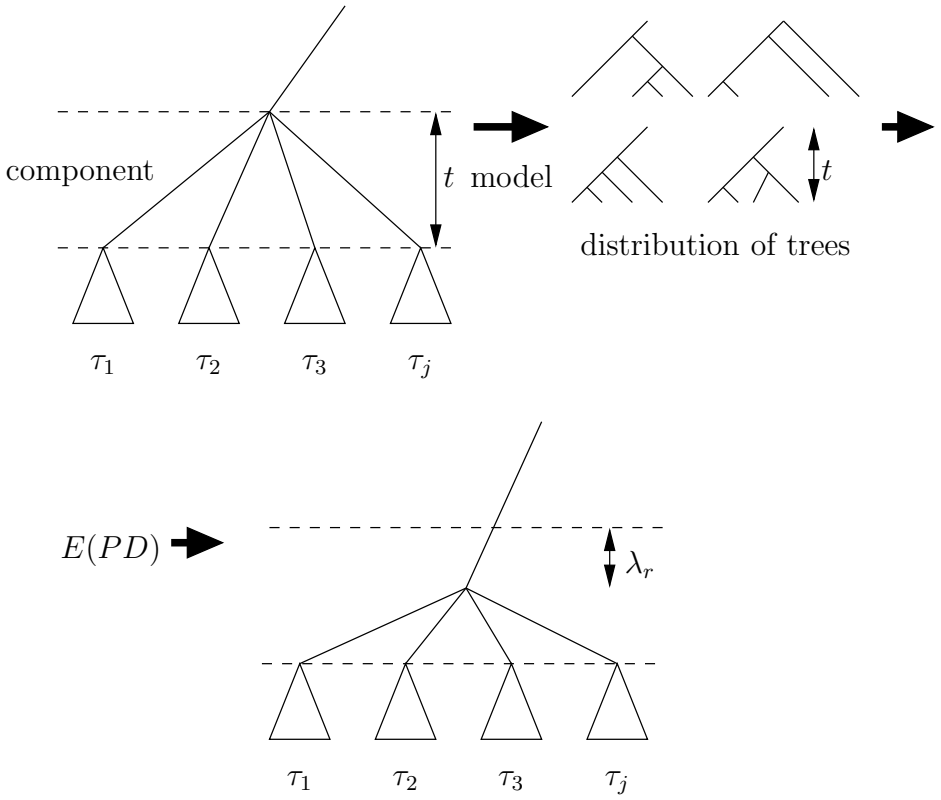


Figure 11.5: The evolutionary model gives a distribution of trees for the component with the polytomy. These trees have an expected PD which is used to rescale the polytomy component.

therefore:

$$\begin{aligned} \bar{PD} &= E(PD|j+1\text{th speciation at } t) \\ &= \sum_{i=2}^{j} iE(s_i|j+1\text{th speciation at } t) \end{aligned}$$

There are many ways of replacing the component of the tree with one with the correct PD. However there is a unique replacement that retains the distances of species from the root of the tree. This replacement has a root edge of length:

$$\lambda_r = \frac{tj - \bar{PD}}{j-1}$$

and 'pendant' edges of length $t - \lambda_r$ (see Figure 11.5).

This procedure can be applied throughout the tree to each polytomy. For the current definition of t it must be applied top down (otherwise rescaling an edge for a polytomy that is below another polytomy may effect the value of t for the latter). The approaches in chapters 2 and 3 and Gernhard (2008) and Gernhard (2007) can be adapted to obtain the required probability densities or tree samples.

11.5 How can unsampled species be included?

Consider the phylogeny available for a given conservation problem. In many situations this phylogeny may not include all the species of interest as the required information (eg. genetic sequences) may be missing for some species. Some information about the approximate location of the missing species in the tree may be available from existing taxonomies or expert opinion. The problem is to produce a species prioritisation that utilises the available phylogeny and any information regarding the approximate position of the missing species in the tree.

The simplest approach would be to ignore the missing species. Without further knowledge of their position in the tree the missing species will generally have a very average expected biodiversity contribution. They are therefore unlikely to be selected for conservation unless a high proportion of species are conserved or they are endangered but cost-effective to conserve.

The problem with ignoring missing species is that they may not be uniformly distributed throughout the tree. Certain clades of particular scientific interest may be fully sampled whilst other clades of lesser interest (or for which collecting specimens is more difficult) may contain more missing species. Simply ignoring the number of missing species will therefore result in a bias that is either towards or against undersampled clades. Undersampled clades will contain less PD than they really do and may therefore be deemed of less importance. If survival probabilities are taken into account the undersampled clades will be at greater risk of complete extinction than they really are and may therefore receive a higher priority.

To address the bias due to missing species we need to use whatever information is available about the location of the missing species in order to

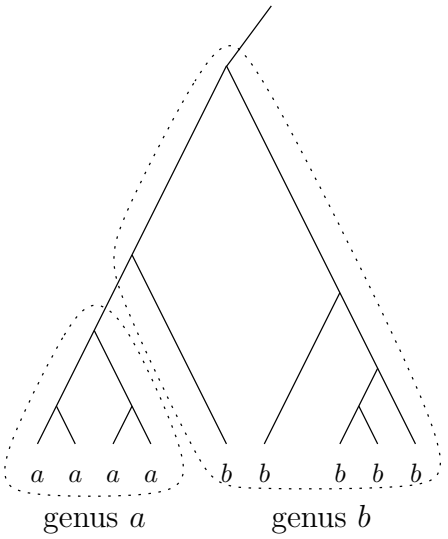


Figure 11.6: This figure shows a portion of a tree containing two genera, one which is actually contained within the other. A missing species of genus a or b would be permitted to connect anywhere within the appropriate circled genera.

i) boost the PD of undersampled clades to their expected values ii) determine their appropriate extinction probabilities. I have done some preliminary work in this area which is still ongoing. My approach has been to generate a sample of trees with missing species included and then apply a prioritisation method to each sampled tree.

To generate the sample of trees MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) was modified to allow non-monophyletic constraints. The input file includes the sequences for the species where available and completely missing data for the missing species. A set of constraints is then specified that forces missing species to be in the 'known' portions of the tree. This approach constructs the tree and adds the missing species in a single step, permitting constraints as illustrated in Figure 11.6.

The output from MrBayes is a large sample of trees. The chosen prioritisation approach is then applied to each sampled tree individually and these prioritisations compared to find, for example, the proportion of sampled trees for which each species is in an optimal set of size k. Alternatively a splits

network could be constructed from the sampled trees and species identified using methods such as those in Spillner et al.; Minh et al..

This work was motivated by an application to the global bird prioritisation for the EDGE of Existence project (Zoological Society of London, 2008). This project seeks to prioritise species on a global basis. The project I considered with Arne Mooers, David Redding and Walter Jetz was the prioritisation of the approximately 10,000 global species of birds. The approach chosen by the EDGE project is to calculate the FP index for each species and simply multiply these by the species' extinction risk, a value they refer to as the EDGE statistic. The aim is to identify species with the highest 100 EDGE values. For the bird tree approximately half of these species are missing, however from their taxonomic classification their rough location in the tree is known. Consequently the method described here will be applied to produce a final prioritisation. I anticipate this to be one of the most influential outcomes from this thesis.

Bibliography

Aldous, D. and L. Popovic. 2005. A critical branching process model for biodiversity. Advances in Applied Probability 37:1094–1115.

Aldous, D. J. 1996. Probability distributions on cladograms. Pages 1–18 *in* Random discrete structures (D. J. Aldous and R. Pemantle, eds.) no. 76 in IMA Volumes Math. Appl. Springer.

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statistical Science 16:23–34.

Altschul, S. F. and D. J. Lipman. 1990. Equal animals. Nature 348:493–494.

Barker, G. M. 2002. Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. Biological Journal of the Linnean Society 76:165–194.

Bertelli, S. and N. P. Giannini. 2005. A phylogeny of extant penguins (Aves: Spenisciformes) combining morphology and mitochondrial sequences. Cladistics 21:209–239.

Bininda-Edmonds, O. R. P., M. Cardillo, K. E. Jones, R. D. E. MacPhee, R. M. D. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman, and A. Purvis. 2007. The delayed rise of present-day mammals. Nature 446:507–512.

Blum, M. and O. Francois. 2006. Which random processes describe the Tree of Life? A large-scale study of phylogenetic tree imbalance. Systematic Biology 55:685–691.

Bokal, D. Computing quadratic entropy for evolutionary trees. in preparation .

Bunnell, F. L. and D. J. Huggard. 1999. Biodiversity across spatial and temporal scales: problems and opportunities. Forest Ecology and Management 115:113–126.

Camm, J. D., S. K. Norman, S. Polasky, and A. R. Solow. 2006. Nature reserve site selection to maximize expected species covered. Operations Research 50:946–955.

Chan, K. M. A. and B. R. Moore. 1999. Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry. American Naturalist 153:332–346.

Chan, K. M. A. and B. R. Moore. 2002. Whole-tree methods for detecting differential diversification rates. Systematic Biology 51:855–865.

Clarke, K. and R. Warwick. 1998. A taxonomic distinctness index and its statistical properties. Journal of Applied Ecology 35:523–531.

Colless, D. H. 1982. Review of: Phylogenetics: The theory and practice of phylogenetic systematics. Systematic Zoology 31:100–104.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2002. Introduction to algorithms. 2nd ed. MIT Press.

Crozier, R. H. 1992. Genetic diversity and the agony of choice. Biological Conservation 61:11–15.

Crozier, R. H. 1997. Preserving the information content of species: Genetic diversity, phylogeny, and conservation worth. Annual Review of Ecology and Systematics 28:243–268.

Crozier, R. H., P. Agapow, and L. J. Dunnett. 2006. Conceptual issues in phylogeny and conservation: a reply to Faith and Baker. Evolutionary Bioinformatics Online 2:197–199.

Crozier, R. H., L. J. Dunnett, and P. M. Agapow. 2005. Phylogenetic biodiversity assessment based on systematic nomenclature. Evolutionary Bioinformatics Online 1:11–36.

Crump, K. S. and C. J. Mode. 1968. A general age-dependent branching process I. Journal of Mathematical Analysis and Applications 24:494–508.

Crump, K. S. and C. J. Mode. 1969. A general age-dependent branching process II. Journal of Mathematical Analysis and Applications 25:8–17.

Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN, a database of homologous vertebrate genes. Nucleic Acids Research 22:2360–2365.

Faith, D., G. Carter, G. Cassis, S. Ferrier, and L. Wilkie. 2003. Complementarity, biodiversity viability analysis, and policy-based algorithms for conservation. Environmental science & Policy 6:311–328.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10.

Faith, D. P. and A. M. Baker. 2006. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evolutionary Bioinformatics Online 2:70–77.

Faith, D. P. and K. J. Williams. 2006. Phylogenetic diversity and biodiversity conservation. Pages 233–235 *in* McGraw-Hill Yearbook of Science and Technology.

Faller, B., F. Pardi, and M. Steel. 2008. Distribution of phylogenetic diversity under random extinction. Journal of theoretical biology 251:286–296.

Garey, M. R. and D. S. Johnson. 1979. Computers and Intractability. W. H. Freemand and Company, San Francisco.

Gaston, K. J. 1996. Species richness: measure and measurement. Pages 77–113 *in* Biodiversity: a biology of numbers and difference (K. Gaston, ed.). Blackwell Science, Cambridge.

Gernhard, T. 2006. Stochastic models of speciation events in phylogenetic trees. Diplom thesis, Technical University of Munich, Germany .

Gernhard, T. 2007. New analytic results for speciation times in neutral models. Submitted .

Gernhard, T. 2008. The conditioned reconstructed process. Journal of theoretical biology, in press .

Gernhard, T., D. Ford, R. Vos, and M. Steel. 2006. Estimating the relative order of speciation or coalescence events on a given phylogeny. Evolutionary Bioinformatics Online 2:309–317.

Gernhard, T., K. Hartmann, and M. Steel. Probability distributions on generalised Yule trees with biodiversity applications. Journal of Mathematical Biology, submitted .

Giannini, N. P. and S. Bertelli. 2004. Phylogeny of extant penguins based on integumentary and breeding characters. The Auk 121:422–434.

Haake, C., A. Kashiwada, and F. E. Su. 2008. The shapley value of phylogenetic trees. Journal of Mathematical Biology 56:479–497.

Hahn, M. W., T. De Bie, J. E. Stajich, C. Nguyen, and N. Cristianini. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res 15:1153–1160.

Harding, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Advances in Applied Probability 3:44–47.

Hartmann, K. 2007. TreeSample. http://www.klaashartmann.com/treesample/.

Hartmann, K., T. Gernhard, and D. Wong. a. Sampling trees from evolutionary models. Systematic Biology, submitted .

Hartmann, K. and A. O. Mooers. When should phylogenies guide conservation management? In preparation .

Hartmann, K., A. O. Mooers, and S. Rioux-Paquette. b. Species conservation with uncertain survival probabilities, in preparation .

Hartmann, K. and M. Steel. 2006. Maximizing phylogenetic diversity in biodiversity conservation: Greedy solutions to the Noah's Ark Problem. Systematic Biology 55:644–651.

Hartmann, K. and M. Steel. 2007. Phylogenetic diversity: from combinatorics to ecology. *in* Reconstructing Evolution - New Mathematical and Computational Advances (O. Gascuel and M. Steel, eds.). Oxford University Press.

Hartmann, K., M. Will, and M. Steel. c. Phylogenetic tree shape as predicted by a lifetime speciation model, in preparation .

Harvey, P. H., R. M. May, and S. Nee. 1994. Phylogenies without fossils. Evolution 48:523–529.

Heard, S. B. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution 50:2141–2148.

Heard, S. B. and A. O. Mooers. 2000. Phylogenetically patterned speciation rates and extinction risks change the loss of evolutionary history during extinctions. Proc. R. Soc. Lond. B 267:613–620.

Heard, S. B. and A. O. Mooers. 2002. Signatures of Random and Selective Mass Extinctions in Phylogenetic Tree Balance. Systematic Biology 51:889–897.

Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46:627–640.

Hohl, M. and M. A. Ragan. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Systematic Biology 56:206–221.

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

Isaac, N. J. B., S. T. Turvey, B. Collen, C. Waterman, and J. E. M. Baillie. 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS One 2:e296.

IUCN. 2007. 2007 IUCN Red list of threatened species. http://www.iucnredlist.org.

Jagers, P. 1975. Branching processes with biological applications. Wiley, New York.

Johst, K., M. Drechsler, and F. Wätzold. 2002. An ecological-economic modelling procedure to design compensation payments for the efficient spatio-temporal allocation of species protection measures. Ecological Economics 41:37–49.

Karev, G. P., Y. I. Wolf, and E. V. Koonin. 2003. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? Bioinformatics 19:1889–1900.

Kingman, J. F. C. 1982a. The coalescent. Stochastic Processes and Their Applications 13:235–248.

Kingman, J. F. C. 1982b. Exchangeability and the evolution of large populations. Exchangeability in Probability and Statistics Pages 97–112.

Kingman, J. F. C. 1982c. On the genealogy of large populations. J. Appl. Probab. 19A:27–43.

Kirkpatrick, M. and M. Slatkin. 1993. Searching for evolutionary pattern in the shape of a phylogenetic tree. Evolution 47:1171–1181.

Korte, B., L. Lovaśz, and R. Shrader. 1991. Greedoids, algorithms and combinatorics. Springer Berlin.

Lamberson, R. H., R. McKelvey, B. R. Noon, and C. Voss. 1992. A dynamic analysis of northern spotted owl viability in a fragmented forest landscape. Conservation Biology 6:505–512.

Lawton, J. H. and R. M. May. 1995. Extinction rates. Oxford University Press, USA.

Lewis, C. A., N. P. Lester, A. D. Bradshaw, J. E. Fitzgibbon, K. Fuller, L. Hakanson, and C. Richards. 1996. Considerations of scale in habitat conservation and restoration. Canadian Journal of Fisheries and Aquatic Sciences 53:440–445.

Lewis, L. A. and P. O. Lewis. 2005. Unearthing the molecular phylodiversity of desert soil green algae (Chlorophyta). Systematic Biology 54:936–947.

Louis Harris & Associaties. 1998. Biodiversity in the Next Millenium Survey. American Museum of Natural History.

Mace, G. M. and R. Lande. 1991. Assessing extinction threats: toward a reevaluation of IUCN categories. Conservation Biology 5:148–157.

Maddison, D. R. and K. S. Schulz. 2006. The tree of life web project. http://tolweb.org.

Magallon, S. and M. J. Sanderson. 2001. Absolute diversification rates in angiosperm clades. Evolution 55:1762–1780.

Matsen, F. A. 2006. A geometric approach to tree shape statistics. Systematic Biology 55:652–661.

McKenzie, A. and M. Steel. 2000. Distributions of cherries for two models of trees. Mathematical Biosciences 164:81–92.

McPeek, M. A. and J. M. Brown. 2007. Clade Age and Not Diversification Rate Explains Species Richness among Animal Taxa. The American Naturalist 169:E97–E106.

Minh, B. Q., S. Klaere, and A. von Haesler. Phylogenetic diversity on split systems. In preparation .

Mooers, A., L. Harmon, M. Blum, D. Wong, and S. Heard. 2007. Some models of phylogenetic tree shape. *in* Reconstructing Evolution - New Mathematical and Computational Advances (O. Gascuel and M. Steel, eds.). Oxford University Press.

Mooers, A. O., D. P. Faith, and W. P. Maddison. Care is needed when converting endangered species lists to extinction probabilities for phylogenetic conservation. In Preparation .

Mooers, A. Ø. and S. B. Heard. 1997. Inferring Evolutionary Process from Phylogenetic Tree Shape. The Quarterly Review of Biology 72:31–54.

Mooers, A. O., S. B. Heard, and E. Chrostowski. 2005. Evolutionary heritage as a metric for conservation. Pages 120–138 *in* Phylogeny and conservation (A. Purvis, T. Brooks, and J. Gittleman, eds.). Cambridge University Press.

Moran, P. A. P. 1958. A general theory of the distribution of gene frequencies. 1. Overlapping generations. Proceedings of the Royal Society of London Series B – Biological Sciences 149:102–112.

Moulton, V., C. Semple, and M. Steel. 2007. Optimizing phylogenetic diversity under constraints. Journal of theoretical biology 246:186–194.

Nee, S., E. C. Holmes, and R. May. 1994. Extinction rates can be estimated from molecular phylogenies. Philosophical transactions of the Royal Society of London B 344:77–82.

Nee, S. and R. M. May. 1997. Extinction and the loss of evolutionary history. Science 278:692–694.

Nee, S., A. O. Mooers, and P. H. Harvey. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proceedings of the National Academy of Sciences of the United States of America 89:8322–8326.

Nelson, W. 1982. Applied Life Data Analysis. Wiley, New York.

Norton, B. G. 1987. Why preserve natural variety? Princeton University Press.

Oakley, T. H., B. Østman, and A. C. V. Wilson. 2006. Repression and loss of gene expression outpaces activation and gain in recently duplicated fly genes. Proceedings of the National Academy of Sciences 103:11637–11641.

Pardi, F. and N. Goldman. 2005. Species choice for comparative genomics: no need for cooperation. PLoS Genetics 1:e71.

Pauplin, Y. 2000. Direct calculation of a tree length using a distance matrix. Journal of Molecular Evolution 51:41–47.

Pavoine, S., S. Ollier, and A. Dufour. 2005. Is the originality of a species measurable? Ecology Letters 8:579–586.

Petchey, O. L. and K. J. Gaston. 2002. Extinction and the loss of functional diversity. Proceedings of the Royal Society of London B 269:1721–1727.

Pimm, S. L., G. J. Russell, J. L. Gittleman, and T. M. Brooks. 1995. The future of biodiversity. Science 269:347–350.

Pinelis, I. 2003. Evolutionary models of phylogenetic trees. Proc. R. Soc. Lond. B 270:1425–1431.

Popovic, L. 2004. Asymptotic genealogy of a critical branching process. Annals of Applied Probability 14:2120–2148.

Pullin, A. S. 2002. Conservation Biology. Cambridge University Press, New York.

Pybus, O. and P. Harvey. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proceedings of the Royal Society of London Series B-Biological Sciences 267:2267–2272.

Rambaut, A. 2002. PhyloGen: Phylogenetic tree simulator package. Department of Zoology, University of Oxford .

Rao, C. P. 1982. Diversity and similarity coefficients: a unified approach. Theoretical Population Biology 21:24–43.

Redding, D., K. Hartmann, A. Mimoto, D. Bokal, M. DeVos, and A. O. Mooers. 2008. The most "original species" often capture more phylogenetic diversity than expected. Journal of theoretical biology, in press 251:606–615.

Redding, D. W. and A. O. Mooers. 2006. Incorporating evolutionary measures into conservation prioritization. Conservation Biology, In Press .

Reist-Marti, S., A. Abdulai, and H. Simianer. 2006. Optimum allocation of conservation funds and choice of conservation programs for a set of African cattle breeds. Genetics Selection Evolution 38:99–126.

Ricklefs, R. E. 2003. Global diversification rates of passerine birds. Proceedings of the Royal Society of London B. 270:2285–2291.

Rodrigues, A. S. L., T. M. Brooks, and K. J. Gaston. 2005. Integrating phylogenetic diversity in the selection of priority areas for conservation: does it make a difference? chap. 5, Pages 101–119 *in* Phylogeny and Conservation (A. Purivs, J. L. Gittleman, and T. Brooks, eds.) no. 8 in Conservation Biology. Cambridge University Press.

Ronquist, F. and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Sanderson, M., M. J. Donoghue, W. Piel, and T. Eriksson. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany 81:183–189.

Sankaranarayanan, G. 1989. Branching processes and its estimation theory. John Wiley and Sons.

Sechrest, W., T. M. Brooks, G. A. B. da Fonseca, W. R. Konstant, R. A. Mittermeier, A. Purvis, A. B. Rylands, and J. L. Gittleman. 2002. Hotspots and the conservation of evolutionary history. Proceedings of the National Academy of Sciences 99:2067–2071.

Secretariat of the Convention on Biological Diversity. 2006. Global biodiversity outlook 2. Montreal.

Semple, C. and M. Steel. 2003. Phylogenetics. Oxford University Press, New York.

Semple, C. and M. Steel. 2004. Cyclic permutations and evolutionary trees. Advances in Applied Mathematics 32:669–680.

Sepkoski, J. J. 1982. Mass extinctions in the Phanerozoic oceans: a review. Geological Society of America Special Paper 190:281–289.

Shaw, A., C. Cox, B. Goffinet, W. Buck, and S. Boles. 2003. Phylogenetic evidence of a rapid radiation of pleurocarpous mosses (bryophyta). Evolution 57:2226–2241.

Simianer, H., S. B. Marti, J. Gibson, O. Hanotte, and J. E. O. Rege. 2003. An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds. Ecological Economics 45:377–392.

Soutullo, A., S. Dodsworth, S. B. Heard, and A. O. Mooers. 2005. Distribution and correlates of carnivore phylogenetic diversity across the Americas. Animal Conservation 8:249–258.

Spillner, A., B. Nguyen, and V. Moulton. Computing phylogenetic diversity for split systems. IEEE/ACM Computational Biology and Bioinformatics, in press .

Steel, M. 2005. Phylogenetic diversity and the greedy algorithm. Systematic Biology 54:527–529.

Steel, M. 2006. Tools to construct and study big trees: A mathematical perspective. *in* Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa (T. R. Hodkinson and J. A. Parnell, eds.). CRC Press.

Steel, M. and A. McKenzie. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Mathematical Biosciences 170:91–112.

Steel, M., A. Mimoto, and A. O. Mooers. 2007. Hedging our bets: the expected contribution of species to future phylogenetic diversity. Evolutionary Bioinformatics 3:237–234.

van der Heide, C. M., J. C. J. M. van den Bergh, and E. C. van Ierland. 2005. Extending Weitzman's economic ranking of biodiversity protection: combining ecological and genetic considerations. Ecological Economics 55:218–223.

Vane-Wright, R. I., C. J. Humphries, and P. H. Williams. 1991. What to protect? - Systematics and the agony of choice. Biological Conservation 55:235–254.

Venditti, C., A. Meade, and M. Pagel. 2006. Detecting the node-density artifact in phylogeny reconstruction. Systematic Biology 55:637–643.

Walters, C. J. 1986. Adaptive management of renewable resources. Macmillan, New York.

Weir, J. T. 2006. Divergent timing and patterns of species accumulation in lowland and highland neotropical birds. Evolution 60:842–855.

Weitzman, M. L. 1992. On diversity. The Quarterly Journal of Economics 107:363–405.

Weitzman, M. L. 1995. Diversity functions. *in* Biodiversity loss: economic and ecological issues (C. Perrings, K.-G. Mäler, C. Folk, C. Holling, and B.-O. Jansson, eds.). Cambridge University Press.

Weitzman, M. L. 1998. The Noah's Ark Problem. Econometrica 66:1279–1298.

Whelan, S., P. I. W. de Bakker, E. Quevillon, N. Rodriguez, and N. Goldman. 2006. Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees (http://www.ebi.ac.uk/goldman-srv/pandit.). Nucleic Acids Research 34:D327–D331.

Wilson, K. A., M. F. McBride, M. Bode, and H. Possingham. 2006. Prioritizing global conservation efforts. Nature 440:337–340.

Witting, L. and V. Loeschcke. 1995. The optimization of biodiversity conservation. Biological Conservation 71:205–207.

Witting, L., J. Tomiuk, and V. Loeschke. 2000. Modelling the optimal conservation of interacting species. Ecological Modelling 125:123–143.

Yule, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Wllis, F.R.S. Philosophical Transactions of the Royal Society of London Series B 213:21–87.

Zoological Society of London. 2008. EDGE of existence. http://www.edgeofexistence.org.

Zwickl, D. J. and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 51:588–598.